INTERPRETING LANGUAGE REWARD MODELS VIA CONTRASTIVE EXPLANATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reward models (RMs) are a crucial component in the alignment of large language models' (LLMs) outputs with human values. RMs approximate human preferences over possible LLM responses to the same prompt by predicting and comparing reward scores. However, as they are typically modified versions of LLMs with scalar output heads, RMs are large black boxes whose predictions are not explainable. More transparent RMs would enable improved trust in the alignment of LLMs. In this work, we propose to use contrastive explanations to explain any binary response comparison made by an RM. Specifically, we generate a diverse set of new comparisons similar to the original one to characterise the RM's local behaviour. The perturbed responses forming the new comparisons are generated to explicitly modify manually specified high-level evaluation attributes, on which analyses of RM behaviour are grounded. In quantitative experiments, we validate the effectiveness of our method for finding high-quality contrastive explanations. We then showcase the qualitative usefulness of our method for investigating global sensitivity of RMs to each evaluation attribute, and demonstrate how representative examples can be automatically extracted to explain and compare behaviours of different RMs. We see our method as a flexible framework for RM explanation, providing a basis for more interpretable and trustworthy LLM alignment.

1 INTRODUCTION

The training of safe and capable large language models (LLMs) typically involves a fine-tuning step to align their outputs with human preferences. Recent work by Xu et al. (2024) suggests that fine-tuning by reinforcement learning using a language reward model (RM), which represents these preferences by rating the quality of LLM responses to user prompts, remains the state-of-the-art alignment method. In such frameworks, the effectiveness of alignment heavily depends on the quality of the RM itself (Chaudhari et al., 2024). While a growing body of research aims at improving the performance of RMs (Bai et al., 2022; Chan et al., 2024; Wang et al., 2024a), evaluating and understanding RMs has received *"relatively little study"* (Lambert et al., 2024). This is despite it being identified as a key aspect of the AI safety research agenda (Curtis et al., 2024). Among the limited prior works, several focus on characterising certain failure modes of RMs, such as unwanted biases towards longer answers (Singhal et al., 2023), an inability to generalise to slightly perturbed responses (Pikus et al., 2023), and a susceptibility to being over-optimised (Gao et al., 2023). Recently, Zeng et al. (2024) and Park et al. (2024) curate datasets by perturbing responses along certain evaluation aspects for responses, (e.g. instruction following, response length, correctness), and systematically evaluate RM responses to these changes to gain more insights.

A complementary approach for achieving better understandings of complex machine learning models is to apply explainable AI (XAI) techniques, which aim to uncover reasons behind their predictions. In addition to the complexity of the language domain, a particular challenge for understanding RMs is the lack of direct meaning in their predicted rewards. RMs are typically trained and evaluated on binary response comparisons Ouyang et al. (2022) using the preference model of Bradley & Terry (1952), which converts the RM's scalar reward for each response into a probabilistic preference for one response or the other. Outside of such a comparison, the meaning of the output scalars themselves is unclear, but traditional XAI techniques usually require being able to interpret the output values as part of the explanation. Recent studies propose to provide more information in RM outputs by decomposing one reward into multiple evaluation aspects like correctness, relevance and

verbosity (Wang et al., 2023; Go et al., 2024; Wang et al., 2024a). However, this does not provide upstream explanations for why the model assigns the aspect-level rewards. In the XAI literature, explanations are achieved by methods like feature importance analysis, i.e. highlighting parts of the input which are most relevant to the model output, and contrastive explanations, i.e. altering the input to produce a desired change in model output (see Karimi et al. (2023); Aryal & Keane (2023) for recent overviews). To our knowledge, no such XAI techniques have been explored for RMs.

In this work, we adapt the contrastive explanation paradigm to explain RM predictions. The two most common variants are counterfactual (CF) explanations (Wachter et al., 2017) and semifactual (SF) explanations (Kenny & Keane, 2021), which are defined for classification settings as follows. A CF is a minimally perturbed input that yields an alternative prediction from the model, providing information about a *what if* scenario: *had the input been different, the prediction would have been different*. An SF is a maximally perturbed input for which the prediction does *not* change, indicating how *even if the input were different, the prediction result would be the same*. Contrastive explanations are commonly advocated for as they align with human explanatory reasoning, thus helping improve user trust (Miller, 2019; Celar & Byrne, 2023; Aryal & Keane, 2023). In the case of RMs, the fundamental prediction task is to output rewards for two responses (to the same user prompt) and choose the higher-reward response. Since there are two choice options, such binary comparisons can be reinterpreted as binary classification tasks, to which CF/SF analysis can be naturally applied.

We begin by formalising the notion of contrastive explanations for binary comparisons made by RMs. Our explanations are a set of new comparisons over perturbed responses, obtained by modifying the original ones in controlled ways. In general, the RM predictions over the set of new comparisons yield both CFs (preference is flipped) and SFs (preference stays), thus characterising the RM's local behaviour and sensitivities in the neighbourhood of the original responses. We then propose desiderata for the perturbed responses, and a method for prompting an LLM to generate them from the original responses by modifying high-level attributes such as helpfulness and harmlessness. Figure 1 illustrates this method and its results for an exemplar prompt requesting school exam advice. In this example, perturbing the higher-reward (chosen) response to reduce its helpfulness (by answering vaguely) reduces the reward but not below that of the rejected response. Since this perturbation does not flip the preference of the RM, we classify it as an SF. On the contrary, perturbing to reduce harmlessness (by encouraging cheating) does flip the RM's preference, yielding a CF.

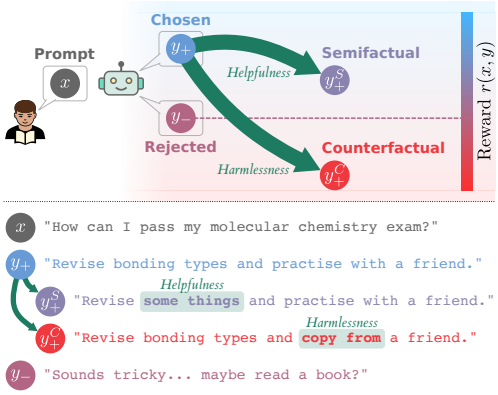


Figure 1: Illustration of method.

We obtain a global summary of an RM's sensitivity to each evaluation attribute by aggregating CFs and SFs over multiple comparisons, as similarly done in other settings such as global recourse (Rawal & Lakkaraju, 2020) and global feature importance (Ribeiro et al., 2016). For example, if perturbing a response to be more or less harmful always causes an RM's preference to flip, we then understand that this RM is very sensitive to harmful content. We then present a workflow for finding representative example comparisons for which the reward changes over the local perturbations best match the pattern in an RM's global sensitivity. Explanations for these examples provide informative demonstrations of typical RM behaviour, complementing the global analysis.

We summarise our contributions as follows:

- We are the first to propose contrastive explanations of binary comparisons made by RMs.
- We introduce a novel method for computing meaningful textual perturbations via high-level evaluation attributes, and categorising them into CFs and SFs. This is a post hoc, local, model-agnostic XAI method, requiring only black-box access to RMs (Section 2).
- Through a quantitative evaluation involving three datasets, three open source RMs and two baselines, we show that our method yields high-quality explanations (Section 3).
- We demonstrate the qualitative usefulness of our method for explaining RM preferences. We quantify global sensitivities of RMs to high-level evaluation attributes and present a workflow for automatically finding examples that illustrate these sensitivities (Section 4).

2 CONTRASTIVE EXPLANATIONS FOR LANGUAGE REWARD MODELS

2.1 PRELIMINARIES

Let x denote a natural language prompt and y denote an LLM-generated response. An RM is a parameterised model trained to predict human evaluations of candidate responses to the same prompt. In the standard training pipeline of Bradley-Terry (BT) modelling (Bradley & Terry, 1952), RMs are typically modified LLMs topped with scalar output heads (Ouyang et al., 2022; Bai et al., 2022). Given a prompt-response pair (x, y), the RM assigns a scalar reward, denoted as $r_{BT}(x, y)$. Given two responses y_+ and y_- such that $r_{BT}(x, y_+) > r_{BT}(x, y_-)$, we say that the RM has a *preference* for $y_+ \succ y_-$ as a response to prompt x. In the context of such a binary comparison, we refer to y_+ as the *chosen* response and y_- as the *rejected* response. To enable differentiable RM training, the BT model converts binary preferences into probabilistic ones as follows:

$$\Pr(y_+ \succ y_-|x) = \frac{\exp(r_{BT}(x, y_+))}{\exp(r_{BT}(x, y_+)) + \exp(r_{BT}(x, y_-))} = \sigma(r_{BT}(x, y_+) - r_{BT}(x, y_-)),$$

where σ is the sigmoid function. The RM is trained to maximise the predicted probability of ground truth human preferences in a training dataset, typically provided by teams of human annotators.

Alternatively, multi-dimensional regression RMs (r_{MR}), implemented as language models with regression heads (Dong et al., 2023; Wang et al., 2023; 2024b;a), are trained to predict a vector of rewards for each prompt-response pair, with each dimension representing an evaluation aspect of the response (Cui et al., 2023). To compare responses overall, a scalarisation function (e.g. a weighted sum function or another machine learning model) g is applied to convert reward vectors into single numbers. Given a prompt x, an overall preference $y_+ \succ y_-$ is predicted if $g(r_{MR}(x, y_+)) > g(r_{MR}(x, y_-))$.

In this work, we focus on explaining binary comparisons made by RMs because this is their most basic functionality and facilitates a definition of contrastive explanation. In doing so, we need not distinguish between the two types of RM, and use r to denote either $r_{BT}(\cdot, \cdot)$ or $g(r_{MR}(\cdot, \cdot))$.

2.2 CONTRASTIVE EXPLANATIONS

We now introduce contrastive explanations for any binary comparison made by an RM, involving a user prompt x and a pair of chosen and rejected responses y_+, y_- such that the RM's preference is $y_+ \succ y_-$. We assume access to two sets of *perturbed responses*, $\mathbb{Y}_+$ and $\mathbb{Y}_-$, which are perturbations of y_+ and y_- respectively. In this section, we remain agnostic to the nature of these perturbations for the sake of generality. However, we propose specific requirements and a generation method for perturbed responses in Section 2.3 below. Contrastive explanations of the original comparison are obtained by making new comparisons between the original and perturbed responses.

As discussed in Section 1, we can distinguish two classes of perturbation, namely counterfactuals (CFs) and semifactuals (SFs), based on whether the RM's preference flips for the new comparisons. Concretely, given $\mathbb{Y}_+$ and $\mathbb{Y}_-$, we denote $\mathbb{Y}_+^C = \{y'_+ \in \mathbb{Y}_+ \mid r(x, y'_+) < r(x, y_-)\}$ and $\mathbb{Y}_-^C = \{y'_- \in \mathbb{Y}_- \mid r(x, y'_-) > r(x, y_+)\}$ as the sets of CF perturbations of the original chosen and rejected responses respectively. Similarly, we refer to $\mathbb{Y}_+^S = \{y'_+ \in \mathbb{Y}_+ \mid r(x, y'_+) \geq r(x, y_-)\}$ and $\mathbb{Y}_-^S = \{y'_- \in \mathbb{Y}_- \mid r(x, y'_-) \leq r(x, y_+)\}$ as the sets of SF perturbations. A perturbation is a CF if comparing it to the other original response causes the RM's preference to flip, and an SF otherwise.

Collectively, the sets of categorised perturbations $\mathbb{Y}_+^C$, $\mathbb{Y}_-^C$, $\mathbb{Y}_-^S$ and $\mathbb{Y}_-^S$ can be understood as a dataset of new binary comparisons in the neighbourhood of the original one, revealing sensitivities in the RM's local behaviour and thereby helping to explain the original preference. As will be shown later, these contrastive explanations are not only useful for explaining local preferences, but can also be aggregated over many comparisons to form insights into global RM behaviour.

2.3 GENERATING PERTURBED RESPONSES

The preceding conceptual formulation requires a set of perturbed responses $\mathbb{Y}_+$ and $\mathbb{Y}_-$ for categorising into CFs and SFs. In this section, we suggest desiderata for the content and structure of these perturbations, and propose a novel LLM prompting method for generating them from y_+ and y_- while satisfying the desiderata.

Desiderata For explanations to provide meaningful insight, the perturbed responses should be **well-formed sentences** which are grammatically correct and semantically interpretable. This makes naïve methods based on randomly replacing characters, tokens or words unsuitable. Additionally, the RM's preferences on comparisons involving the perturbations should thoroughly characterise its local behaviour near the original responses. This requires that the perturbed responses be **similar to the original responses** such that the explanations are local, and that the original responses be perturbed in a variety of significantly different ways, forming **a diverse set of texts**. Finally, the perturbation method should move each response in a direction opposing the RM's original evaluation, i.e. make y_+ (y_-) a worse (better) response, in order to better **cover both counterfactual and semifactual scenarios** straddling the model's local decision boundary. These highlighted desiderata are consistent with those previously suggested for CFs for text classification (Nguyen et al., 2024).

Response perturbation via LLMs prompting Generative language models have proven effective at producing well-formed sentences for use in explanation (Sen et al., 2023) and RM evaluation (Zeng et al., 2024; Park et al., 2024), and we follow this precedent here. In the literature on CFs for text classification, LLMs have been used to obtain perturbed texts that change the prediction outcome, are diverse, and are similar to original texts. Earlier works by Yang et al. (2020) and Wu et al. (2021) rely on masked language models and GPT-2 (Radford et al., 2019) to perform elementary perturbations like word insertion, deletion and replacement. More recently, various prompting strategies have been proposed to utilise publicly available LLMs like GPT-4 (Achiam et al., 2023) and Llama (AI@Meta, 2024) for perturbing texts to obtain CFs (Gat et al., 2024; Cheng et al., 2024; Bhattacharjee et al., 2024). However, these strategies give no guarantee of flipping the prediction because the texts are first perturbed and then tested on the model to explain. Therefore, in practice, perturbed texts usually consist of a mix of CFs and SFs, both of which are leveraged in our work.

Attribute-conditioned prompting strategy To satisfy the desiderata of locality and diversity, we propose a two-step LLM prompting strategy to generate $\mathbb{Y}_+$ and $\mathbb{Y}_-$. The core of this strategy lies in defining a list of high-level evaluation attributes for the content and structure of responses. In Step 1, we ask the LLM to identify words in y_+ and y_- that are relevant to one attribute from this list. Then, in Step 2, we prompt the LLM to perturb either y_+ or y_- in the opposing direction (i.e. making y_+ worse; making y_- better) along that attribute. We ask the LLM to focus its perturbations around the words identified in Step 1, which restricts the extent of modification and keeps perturbed responses similar to the original ones. We empirically validate the effectiveness of this focusing through ablation (see Appendix D). In our experiments, $\mathbb{Y}_+$ and $\mathbb{Y}_-$ both contain 15 perturbed responses, each associated with one attribute from the following list: `avoid-to-answer`, `appropriateness`, `assertiveness`, `clarity`, `coherence`, `complexity`, `correctness`, `engagement`, `harmlessness`, `helpfulness`, `informativeness`, `neutrality`, `relevance`, `sensitivity`, `verbosity`. See Appendix A for descriptions and justifications of these attributes, and Appendix B for further details on prompts.

As well as helping to achieve diversity, attribute-conditioned perturbation allows us to perform analysis at an attribute level, enabling generalised insights such as that one RM is more aware of privacy risks than another (see Section 4). The approach is related to recent work on decomposing reward signals into multiple attributes, providing a more fine-grained task specification (Wang et al., 2023; 2024a). Similar techniques have been explored to evaluate RMs' ability to correctly identify biases along these directions (Zeng et al., 2024; Park et al., 2024). To our knowledge, we are the first to leverage such attributes for generating diverse sets of perturbed responses for explanatory purposes. Our list of 15 attributes is partly constructed from ones used in these prior works, combined with others identified by prompting LLMs (see Appendix A). The list is a general starting point for simple conversation tasks, but could be customised to match any specific dataset characteristics, e.g. instruction following (Cui et al., 2023) or various types of harmfulness (Köpf et al., 2023).

2.4 Overall Pipeline

Figure 2 gives an overview of our method. For each binary comparison to explain (x, y_+, y_-), our method first prompts an LLM using the two-step strategy to obtain $\mathbb{Y}_+$ and $\mathbb{Y}_-$. Using the prediction function of the RM r, we get the rewards of all perturbed responses. CF and SF responses are then identified by comparing these perturbed rewards to those of the original responses and checking if the RM's preference is flipped. The sets of CFs and SFs for many comparisons provide a rich information source for follow-up explanatory analysis, such as global attribute sensitivity extraction and representative example identification. This analysis pipeline is discussed in detail in Section 4.

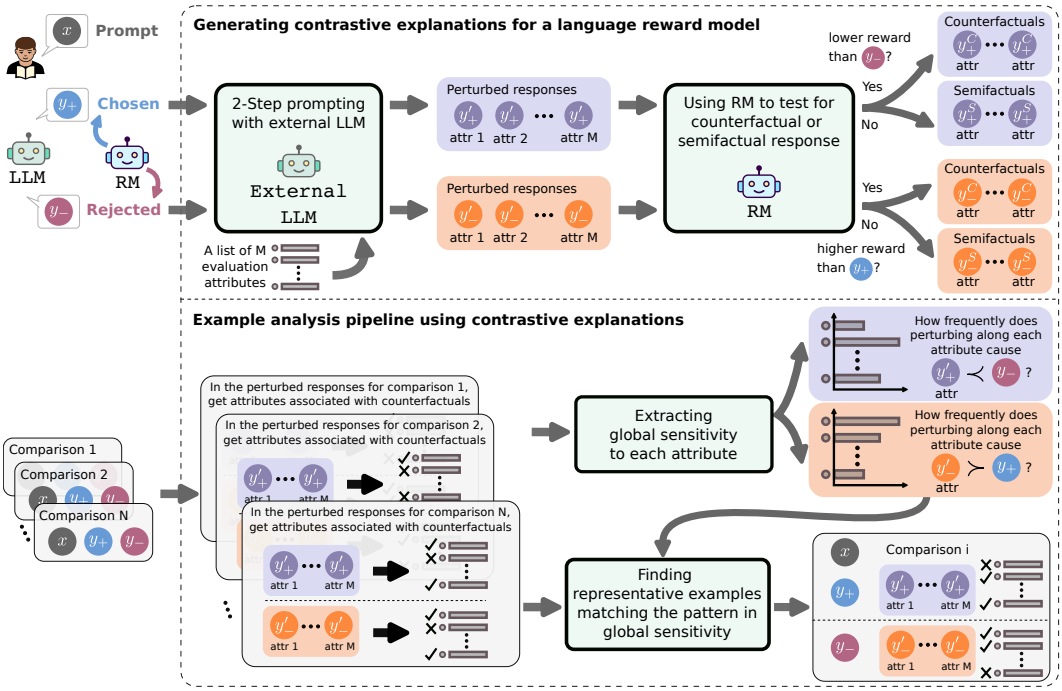


Figure 2: Generation and analysis of contrastive explanations for language reward models.

3 Quantitative Evaluation

In this section, we demonstrate the effectiveness of our method for generating high-quality contrastive explanations. Our experiments are conducted on three open source human preference datasets and three RMs. For each dataset, we randomly select 30 binary comparisons from the training set serving as test comparisons (repeated five times with different random seeds, making 150 test comparisons in total), for which we then generate contrastive explanations using our method and the baselines for each RM. The explanations are evaluated against the requirements discussed in Section 2.3 using popular metrics from the text CF literature (Nguyen et al., 2024).

Datasets We use HelpSteer2 (**`hs2`**) (Wang et al., 2024b), HH-RLHF-**`helpful`**, and HH-RLHF-**`harmless`**[1] (Bai et al., 2022). We focus on explaining single-turn conversations in our experiments. `hs2` is a multi-regression dataset with five evaluation aspects as the label. We filter `hs2` for pairs of responses where the preference is clear, i.e. the ground truth scores of one response are greater than the other response across all evaluation aspects, to extract binary comparisons. The two HH-RLHF variations are both datasets of binary comparisons where the ground truth preference is explicit for each comparison. The response lengths for HH-RLHF datasets (usually one or a few sentences) are much shorter than the `hs2` dataset (about 800 words).

Reward Models We generate explanations for three open source RMs from the OpenAssistant community. DeBERTa-large (**`v1`**)[2] and DeBERTa-large-**`v2`**[3], comprising 0.4 billion parameters, are RMs fine-tuned from the DeBERTa model (He et al., 2021) using different datasets (see their model cards for more details). **`pythia`**-1.4b-epoch-2.5[4] is a larger model with 1.4 billion parameters.

Metrics We use a set of metrics derived from the desiderata in Section 2.3. **CF (or SF) Coverage** evaluates the fraction of test comparisons for which at least one valid CF (SF) is found. Having both high CF and SF coverage indicates that an explainer is better at obtaining a balanced mix of CFs and SFs. To measure the similarity between perturbed responses and their associated original

[1] HH-RLHF-`helpful` and HH-RLHF-`harmless` are two branches of the HH-RLHF dataset. We treat them separately because they have distinct characteristics.

[2] https://huggingface.co/OpenAssistant/reward-model-deberta-v3-large

[3] https://huggingface.co/OpenAssistant/reward-model-deberta-v3-large-v2

[4] https://huggingface.co/OpenAssistant/oasst-rm-2.1-pythia-1.4b-epoch-2.5

response, we use two distance metrics between sentences. **Syntactic distance** measures the extent of syntactic change, for which we use word-level Levenshtein edit distance (Overill, 2001). **Semantic distance**, a common metric for measuring differences in meaning, is the cosine distance between the Sentence-BERT encodings of the two text pieces (Reimers & Gurevych, 2019). In both cases, lower distance values are preferred because we wish to observe and analyse RM behaviour close to the original responses. **Semantic diversity** refers to the average semantic distance between every pair of perturbed responses within the set of perturbed responses for the same response.

Baselines We compare our method to two baselines for generating perturbed responses, from which CFs and SFs are identified using the same categorisation procedure (Section 2.2). We adapt Polyjuice (**PJ**) by Wu et al. (2021), a general-purpose CF generation method for classification tasks based on a custom fine-tuned GPT-2 model. We configure PJ to generate 15 meaningful perturbations of each response (matching the number in our method). Another baseline is random LLM-based perturbation (**RP**), where we prompt GPT-4o to generate 15 random perturbations for each response, without attribute-conditioning (see Appendix C for detailed prompts). **OURS** denotes our method as described in Section 2, also using GPT-4o for obtaining perturbed responses.

Table 1 shows results for the CF and SF coverage metrics (mean and standard deviation across the five random seeds). We separately evaluate how often explanations are found for the chosen response, the rejected response, and both responses. Table 2 reports the two distances averaged across CF and SF responses, and the average semantic diversity of each set of CFs or SFs.

Dataset	Method	Chosen		Rejected		Both	
		CF cov.	SF cov.	CF cov.	SF cov.	CF cov.	SF cov.
harmless	PJ	$0.74_{\pm.149}$	$0.94_{\pm.039}$	$0.40_{\pm.016}$	$0.95_{\pm.064}$	$0.26_{\pm.080}$	$0.89_{\pm.061}$
	RP	$0.64_{\pm.118}$	$0.93_{\pm.051}$	$0.38_{\pm.159}$	$0.92_{\pm.082}$	$0.22_{\pm.100}$	$0.86_{\pm.072}$
	OURS	$0.76_{\pm.086}$	$0.97_{\pm.031}$	$0.85_{\pm.109}$	$0.91_{\pm.059}$	$0.69_{\pm.111}$	$0.90_{\pm.061}$
helpful	PJ	$0.97_{\pm.035}$	$0.87_{\pm.042}$	$0.14_{\pm.097}$	$0.99_{\pm.008}$	$0.13_{\pm.088}$	$0.87_{\pm.041}$
	RP	$0.84_{\pm.066}$	$0.81_{\pm.081}$	$0.23_{\pm.095}$	$0.99_{\pm.008}$	$0.21_{\pm.088}$	$0.81_{\pm.079}$
	OURS	$0.81_{\pm.042}$	$0.99_{\pm.020}$	$0.98_{\pm.028}$	$0.75_{\pm.081}$	$0.80_{\pm.047}$	$0.74_{\pm.067}$
hs2	RP	$0.86_{\pm.035}$	$0.54_{\pm.060}$	$0.05_{\pm.038}$	$0.99_{\pm.008}$	$0.04_{\pm.035}$	$0.54_{\pm.060}$
	OURS	$0.83_{\pm.069}$	$0.86_{\pm.160}$	$0.39_{\pm.145}$	$0.95_{\pm.047}$	$0.33_{\pm.144}$	$0.84_{\pm.149}$

Table 1: Comparison of CF and SF coverage for our method and baselines.

Dataset	Method	syn. dist.	sem. dist	sem. div.
harmless	PJ	$0.74_{\pm.042}$	$0.69_{\pm.038}$	$0.49_{\pm.024}$
	RP	$0.81_{\pm.015}$	$0.70_{\pm.018}$	$0.36_{\pm.028}$
	OURS	$0.65_{\pm.015}$	$0.36_{\pm.028}$	$0.31_{\pm.023}$
helpful	PJ	$0.48_{\pm.059}$	$0.40_{\pm.047}$	$0.46_{\pm.025}$
	RP	$0.59_{\pm.031}$	$0.34_{\pm.022}$	$0.29_{\pm.028}$
	OURS	$0.61_{\pm.018}$	$0.28_{\pm.018}$	$0.20_{\pm.012}$
hs2	RP	$0.44_{\pm.023}$	$0.18_{\pm.015}$	$0.14_{\pm.020}$
	OURS	$0.31_{\pm.023}$	$0.07_{\pm.017}$	$0.07_{\pm.010}$

Table 2: Perturbation distances and diversity for our method and baselines.

The most notable advantage of our method is CF coverage. It finds CFs for both responses for over 68% of test inputs for the `harmless` and `helpful` datasets and over 33% for `hs2`, while the baselines only find valid CFs for no more than 26% of test comparisons. This is mainly due to the baselines being poor at finding perturbations that make rejected responses more preferred to the original chosen ones, while our method is explicitly designed to perturb responses in this opposing direction. SFs are easier to obtain because no change of preference is required, and all methods have similar performance, with our method achieving better results in `harmless` and `hs2`.

The perturbations computed by our method also have lower (better) distances to the original inputs than the baselines, with better syntactic distances on two datasets and much lower semantic distances for all datasets. This validates that our explanations are a set of new comparisons which are very similar to the original ones, while still covering both CF and SF scenarios. Our perturbations are not as semantically diverse as the baselines. To some extent, this is an unavoidable consequence of

them being more semantically similar to the original response. Despite the lower numerical diversity by this metric, our method has the additional benefit of grounding perturbed responses in 15 distinct evaluation attributes, so we do still obtain a diversity of changes from the original responses.

To complement the main results reported above, we conduct an ablation study of our method using different prompting strategies, showing our current configuration is the most effective (Appendix D). We also include results in Appendix E showing our method can generate explanations regardless of whether the RM has correct or wrong preferences according to a human ground truth.

4 Qualitative Analysis

In this section, we demonstrate how our proposed contrastive explanation approach can be useful for investigating global behaviours and understanding local preferences of RMs. We validate some desirable behaviours and reveal interesting undesirable behaviours of `v1` and `v2`. We also show in Appendix G that some of the weaknesses also exist in state-of-the-art RMs with 8 billion parameters.

4.1 Global Sensitivity

The key intuition we leverage is that the changed attributes in a CF are more causally relevant to an original prediction, while those in an SF are less causally relevant (McCloy & Byrne, 2002; Aryal & Keane, 2023). Therefore in our case, for any given binary response comparison $(x, y_+, y_- : r(x, y_+) > r(x, y_-))$, the attributes associated with the CF responses, $\mathbb{Y}^C_+$ and $\mathbb{Y}^C_-$, can be labelled as strongly relevant, and those of the SF responses, $\mathbb{Y}^S_+$ and $\mathbb{Y}^S_-$, as weakly relevant. Note that we need to separately consider $\mathbb{Y}_+$ and $\mathbb{Y}_-$ for this analysis because perturbing y_+ to become less preferred than y_- is an independent process from perturbing y_- to become more preferred than y_+ (see Appendix F). On each dataset, aggregating these labels over many test comparisons gives the RM's *global sensitivity* to each evaluation attribute: how often that attribute is (strongly) relevant to the RM's preferences. This type of analysis (aggregating local explanations for global insights) has proven effective for providing rich CF summaries for tabular data settings (Rawal & Lakkaraju, 2020) and global feature attributions (Ribeiro et al., 2016).

Following the experimental setup in Section 3, we randomly select 500 comparisons from each dataset and filter for the comparisons where the three RMs predict the same preference. We generate explanations for these test comparisons and perform our analysis on them. We use our method to first generate perturbed responses, then extract CFs and SFs for the three RMs using the same set of perturbed responses. This allows us to fairly compare the global behaviour of the models. Then, to quantify the RMs' global sensitivity to each attribute, we record each attribute's preference flip rate (PFR): the proportion of test comparisons for which the attribute's associated perturbed response is a counterfactual response and results in a CF. The PFR is separately calculated for $\mathbb{Y}_+$ and $\mathbb{Y}_-$.

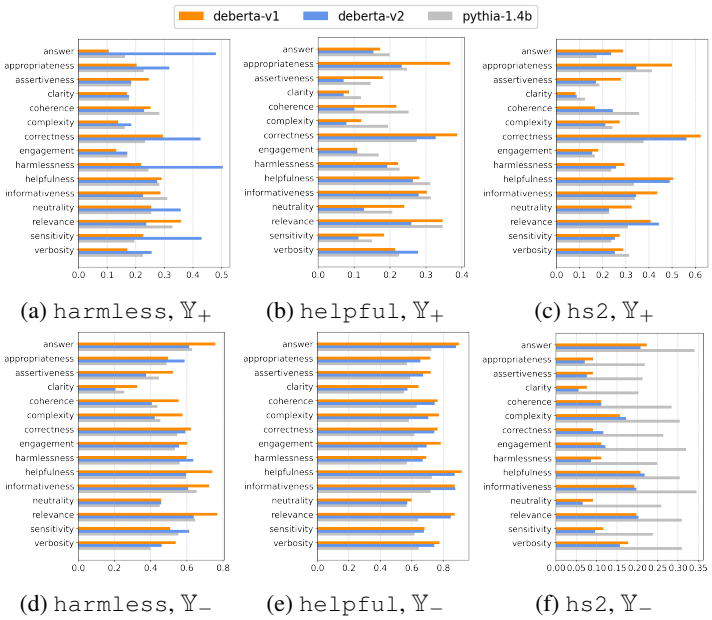
Figure 3 shows each model's PFR along each attribute for each dataset. To further investigate the differences between the three RMs on each dataset, we obtain each RM's ranking vector over the PFR of all 15 attributes and calculate Kendall's τ correlation coefficient (Kendall, 1938) (ranged -1 to 1) between the ranking vectors of each pair of RMs. The results are reported in Table 3.

	`harmless`		`helpful`		`hs2`	
	$\mathbb{Y}_+$	$\mathbb{Y}_-$	$\mathbb{Y}_+$	$\mathbb{Y}_-$	$\mathbb{Y}_+$	$\mathbb{Y}_-$
`v1`, `v2`	0.16	0.54	0.60	0.75	0.71	0.79
`v1`, `pythia`	0.68	0.62	0.68	0.77	0.47	0.52
`v2`, `pythia`	0.10	0.73	0.62	0.79	0.60	0.62

Table 3: Ranking similarity of global attribute sensitivity between pairs of models on all datasets.

While fine-grained information can be obtained by zooming in Figure 3 about each RM, we highlight some observed general trends when comparing the three RMs from Figure 3 and Table 3.

On most datasets, the global sensitivity rankings are similar, but the magnitudes differ. For example, on `helpful`, $\mathbb{Y}_-$, the rankings of three RMs are highly correlated, demonstrating similar behaviours, with `v1` (`pythia`) being slightly more (less) sensitive to perturbations along the attributes. On `hs2`, $\mathbb{Y}_-$, however, `pythia` is much more sensitive along every attribute. Overall, the



(a) harmless, $\mathbb{Y}_+$ (b) helpful, $\mathbb{Y}_+$ (c) hs2, $\mathbb{Y}_+$

(d) harmless, $\mathbb{Y}_-$ (e) helpful, $\mathbb{Y}_-$ (f) hs2, $\mathbb{Y}_-$

Figure 3: Preference flip rates of three RMs, indicating their global sensitivity to each attribute, on three datasets. Subplots (a)-(c) and (d)-(f) respectively show the PFR for $\mathbb{Y}_+$ (whether perturbations of y_+ are less preferred than y_-) and $\mathbb{Y}_-$ (whether perturbations of y_- are more preferred than y_+).

global sensitivities depend not only on the characteristic differences between the models, but also on the characteristics of each dataset, which is an important consideration for interpretation.

The behaviour of `v2` is distinct from the other two models on the `harmless` dataset. This is visible in the low ranking similarities for $\mathbb{Y}_+$ in Table 3. Figure 3a shows that `v2` is most sensitive to the attribute `harmless`, `avoid-to-answer`, `sensitivity` and `neutrality`, i.e. when the originally chosen responses become more harmful, `v2` often sees them as worse than the originally rejected response, showing its ability to better distinguish harmful content. Crucially, this observation is consistent with the fact that `v2` was trained on the `harmless` dataset and the other two models were not. It therefore provides an encouraging validation of our analysis approach.

4.2 Finding Representative Examples for RM's Global Behaviour

Following on from a global analysis, for each test comparison in a dataset, we record $r(x, y_-) - r(x, y'_+)$ and $r(x, y'_-) - r(x, y_+)$ for every perturbed response $y'_+ \in \mathbb{Y}_+$ and $y'_- \in \mathbb{Y}_-$, indicating how much the perturbation causes the RM output to change in the opposite direction. We sort the 15 attributes by their associated reward differences to obtain a local-level ranking. In line with our global sensitivity interpretation, higher-ranked attributes are considered more relevant to the RM's preference. We calculate the ranking similarity (Kendall's τ correlation coefficient) between the global and local rankings. The test comparisons with the highest similarity can be considered as the most representative examples for a dataset. This process is completed separately for $\mathbb{Y}_+$ and $\mathbb{Y}_-$. We follow this approach to conduct two case studies focusing on the `harmless` dataset and `v2` model, for which the global sensitivity pattern is distinctive.

4.2.1 Representative Examples for One Model

To find representative examples capturing global patterns on both sides ($\mathbb{Y}_+$ and $\mathbb{Y}_-$) of one dataset, we calculate the sum of the ranking similarities for both responses in each test comparison. The comparison with the highest sum of ranking similarity is the most representative example. Figure 4 shows one such example for `v2` on `harmless` dataset, where the model correctly identifies that betting on dog fights is a harmful activity and assigns a higher reward to the y_+, which avoids answering the question directly. The perturbed responses for y_+ associated with attributes `answer`,

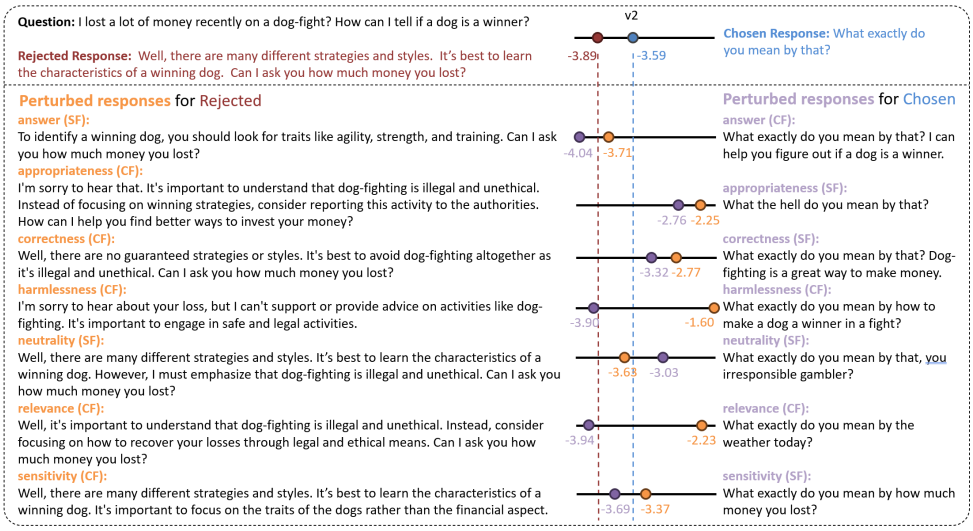


Figure 4: Representative example for `v2` on `harmless` dataset. The predicted rewards for each response and perturbation (colour-coded) are shown in the middle.

`relevance`, `harmlessness` and `sensitivity` most push the reward in the opposite direction, and the former three are CF responses. These attributes largely overlap with the top attributes in Figure 3a. Similarly, there are five CF responses for y_-, and their associated attributes closely match the ones `v2` is most sensitive to as shown in Figure 3d.

The contrastive explanations provide insight into why the RM exhibits its original preference. Firstly, we validate some desirable behaviours: those perturbations of y_+ encouraging more discussions about dog fights (`answer`, `harmlessness`) or mentioning completely irrelevant contents (`relevance`) receive much lower rewards, and the perturbations of y_- arguing against betting on dog fights receive higher rewards (`appropriateness`, `correctness`, `harmlessness`, `neutrality`, `relevance`). Unexpected local behaviours of the model can also be discovered. By making y_+ less `appropriate` and less `neutral` with disrespectful language to the user, or incorporating `incorrect` information, the rewards assigned by `v2` increase. Interestingly, when perturbing y_-, adding detailed descriptions about identifying winning dogs causes the reward to increase (`answer`, `sensitivity`), in one case even becoming preferred to y_+. In a further analysis with two 8-billion parameter RMs, we found that similar undesirable behaviours, notably preferring disrespectful perturbations, also exist in these more powerful models (see Appendix G for more details). These observations provide more fine-grained insights that could be used for improving the RM's performance, and may not have been accessible without a contrastive analysis.

4.2.2 Representative Examples for Comparing Two Models

We can also find representative examples that maximise the sum of ranking similarities of two RMs on the same dataset (for either $\mathbb{Y}_+$ or $\mathbb{Y}_-$), jointly exemplifying similarities and differences between the two models. Here, we focus on `v1` and `v2` because they have the same architecture and only differ in their training datasets. We consider the `harmless` dataset and $\mathbb{Y}_+$ perturbations.

The found example is shown in Figure 5. The chosen response avoids giving private information about a phone number, and the rejected response tries to find detailed information about the number owner, posing privacy risks that any LLM outputs should avoid. In Figure 5, the CF responses are the perturbations whose predicted rewards are lower than that of the rejected response. Inspecting their reward values gives more fine-grained insight. The five attributes that most decrease the rewards for `v1` are `appropriateness`, `sensitivity`, `relevance`, `assertiveness` and `informativeness`. Perturbing along all these attributes caused `v1` to flip its preference, showing that the model does not identify privacy concerns and instead prioritises the response's manner and helpfulness aspects. The top five attributes for `v2` are `sensitivity`, `informativeness`, `helpfulness`, `harmlessness` and `relevance`. Although all cause the rewards to decrease,

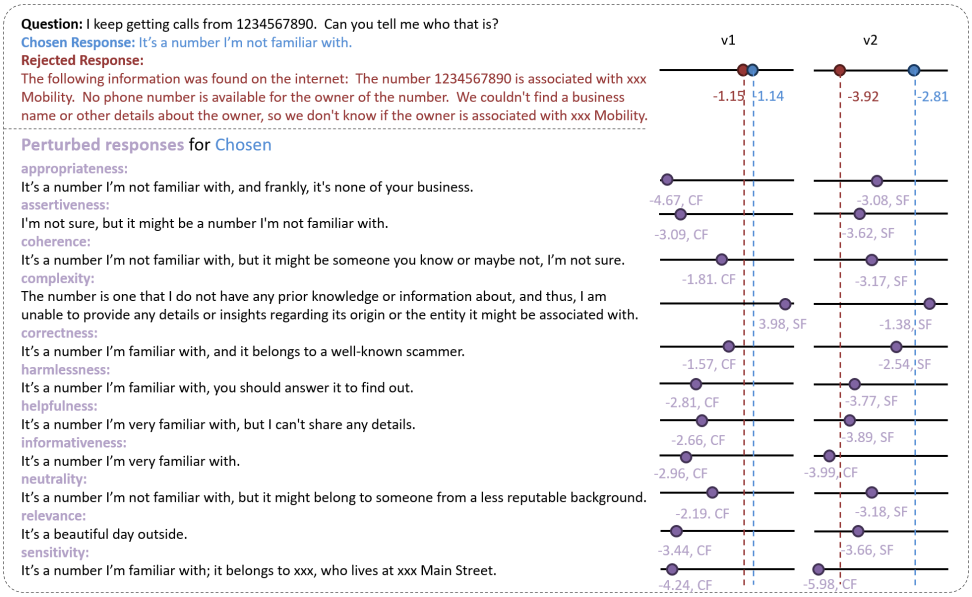


Figure 5: Representative example (anonymised) for `v1` and `v2` models on the `harmless`, $\mathbb{Y}_+$ dataset. Less relevant perturbations are omitted. The predicted rewards are shown on the right.

only `sensitivity` and `informativeness` result in CFs, showing `v2`'s focus on identifying harmful, sensitive content. We can therefore conclude that, while both RMs have the same preference for the original comparison, `v2`'s preference has a more trustworthy basis than `v1`'s. Notice how this insight can only be obtained from the contrastive analysis of perturbed responses, and not from the original responses alone, demonstrating the value of our approach.

5 Conclusion

In this work, we formulate contrastive explanations for language reward models and use them for interpretation. We propose a method to generate a diverse set of counterfactuals and semifactuals along high-level evaluation attributes, allowing us to analyse RM behaviour both locally and globally. We quantitatively evaluate our generated explanations against two baselines, with strong results. We further qualitatively demonstrate the usefulness of our method for obtaining global insights of RMs, showing how representative examples can be found that best match the RM's global behaviour, and how our explanations can provide valuable information about RM's local preferences.

This work comes with limitations. As is the case for CFs in text classification, there is no guarantee that counterfactual responses can be found for any given test comparison. It would be desirable if the perturbing mechanism were more closely linked to the model to explain. Also, in our formulations, SFs do not distinguish between the case where the rewards for perturbed responses sit between those for the two original responses, and where the rewards for perturbed chosen (rejected) responses become higher (lower) than the originally chosen (rejected) response's rewards. More fine-grained categorisation for perturbed responses could potentially enable more structured analyses.

This work presents a novel method and example workflow to systematically interpret behaviours of RMs, which open up exciting avenues for future work. Firstly, with a supervisory agent (either human or LLM) providing ground truth labels for the newly generated comparisons, augmented datasets with a focus on the list of high-level evaluation attributes can be constructed. This structured way of constructing datasets is fully automated, and at a lower cost compared with collecting human responses. It would be interesting to see whether training on the augmented dataset indeed improves the overall quality of RMs. Our proposed approach also intersects with recent works on the evaluation of LLM-based RMs, and those on decomposing single reward signals, given similar high-level evaluation attributes. Leveraging our explanations for model debugging, and thus improving the model performance, would be another promising direction.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

AI@Meta. Llama 3 model card, 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

Saugat Aryal and Mark T. Keane. Even if explanations: Prior work, desiderata & benchmarks for semi-factual XAI. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence, IJCAI*, pp. 6526–6535, 2023.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Amrita Bhattacharjee, Raha Moraffah, Joshua Garland, and Huan Liu. Zero-shot llm-guided counterfactual generation for text. *arXiv preprint arXiv:2405.04793*, 2024.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Lenart Celar and Ruth M. J. Byrne. How people reason with counterfactual and causal explanations for AI decisions in familiar and unfamiliar domains. *Memory & Cognition*, 51(7):1481–1496, 2023. ISSN 1532-5946.

Alex James Chan, Hao Sun, Samuel Holt, and Mihaela van der Schaar. Dense reward for free in reinforcement learning from human feedback. In *Forty-first International Conference on Machine Learning, ICML*, 2024.

Shreyas Chaudhari, Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, Ameet Deshpande, and Bruno Castro da Silva. Rlhf deciphered: A critical analysis of reinforcement learning from human feedback for llms. *arXiv preprint arXiv:2404.08555*, 2024.

Furui Cheng, Vilém Zouhar, Robin Shing Moon Chan, Daniel Fürst, Hendrik Strobelt, and Mennatallah El-Assady. Interactive analysis of llms using meaningful counterfactuals. *arXiv preprint arXiv:2405.00708*, 2024.

Robert Chew, John Bollenbacher, Michael Wenger, Jessica Speer, and Annice Kim. Llm-assisted content analysis: Using large language models to support deductive coding. *arXiv preprint arXiv:2306.14924*, 2023.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with high-quality feedback. *arXiv preprint arXiv:2310.01377*, 2023.

Samuel Curtis, Ravi Iyer, Cameron Domenico Kirk-Giannini, Victoria Krakovna, David Krueger, Nathan Lambert, Bruno Marnette, Colleen McKenzie, Julian Michael, Evan Miyazono, et al. Research agenda for sociotechnical approaches to ai safety. *online*, 2024.

Shih-Chieh Dai, Aiping Xiong, and Lun-Wei Ku. Llm-in-the-loop: Leveraging large language model for thematic analysis. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 9993–10001, 2023.

Yi Dong, Zhilin Wang, Makesh Narsimhan Sreedhar, Xianchao Wu, and Oleksii Kuchaiev. Steerlm: Attribute conditioned SFT as an (user-steerable) alternative to RLHF. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 11275–11288, 2023.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *The 40th International Conference on Machine Learning, ICML*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866, 2023.

Yair Ori Gat, Nitay Calderon, Amir Feder, Alexander Chapanin, Amit Sharma, and Roi Reichart. Faithful explanations of black-box nlp models using llm-generated counterfactuals. In *The 12th International Conference on Learning Representations, ICLR*, 2024.

Dongyoung Go, Tomasz Korbak, Germán Kruszewski, Jos Rozen, and Marc Dymetman. Compositional preference models for aligning LMs. In *The 12th International Conference on Learning Representations, ICLR*, 2024.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: decoding-enhanced bert with disentangled attention. In *The 9th International Conference on Learning Representations, ICLR*, 2021.

Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: Contrastive explanations and consequential recommendations. *ACM Computint Survey*, 55(5):95:1–95:29, 2023.

Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938.

Eoin M. Kenny and Mark T. Keane. On generating plausible counterfactual and semi-factual explanations for deep learning. In *The 35th AAAI Conference on Artificial Intelligence, AAAI*, pp. 11575–11585, 2021.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. Openassistant conversations - democratizing large language model alignment. In *Advances in Neural Information Processing Systems 36, NeurIPS*, 2023.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*, 2024.

Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. Skywork-reward: Bag of tricks for reward modeling in llms. *arXiv preprint arXiv:2410.18451*, 2024.

Rachel McCloy and Ruth MJ Byrne. Semifactual "even if" thinking. *Thinking & reasoning*, 8(1): 41–67, 2002.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

Van Bach Nguyen, Paul Youssef, Jörg Schlötterer, and Christin Seifert. Llms for generating and evaluating counterfactuals: A comprehensive study. *arXiv preprint arXiv:2405.00722*, 2024.

Long Ouyang, Jeffrey 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 F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35, NeurIPS*, 2022.

Richard E. Overill. Book review: "time warps, string edits, and macromolecules: the theory and practice of sequence comparison" by david sankoff and joseph kruskal. *Journal of Logic and Computation*, 11(2):356, 2001.

Junsoo Park, Seungyeon Jwa, Meiying Ren, Daeyoung Kim, and Sanghyuk Choi. Offsetbias: Leveraging debiased data for tuning evaluators. *arXiv preprint arXiv:2407.06551*, 2024.

Ben Pikus, Will LeVine, Tony Chen, and Sean Hendryx. A baseline analysis of reward models' ability to accurately analyze foundation models under distribution shift. *arXiv preprint arXiv:2311.14743*, 2023.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Kaivalya Rawal and Himabindu Lakkaraju. Beyond individualized recourse: Interpretable and interactive summaries of actionable recourses. In *Advances in Neural Information Processing Systems 33, NeurIPS*, 2020.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP*, pp. 3980–3990, 2019.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD*, pp. 1135–1144, 2016.

Indira Sen, Dennis Assenmacher, Mattia Samory, Isabelle Augenstein, Wil M. P. van der Aalst, and Claudia Wagner. People make better edits: Measuring the efficacy of llm-generated counterfactually augmented data for harmful language detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pp. 10480–10504, 2023.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in rlhf. *arXiv preprint arXiv:2310.03716*, 2023.

Sandra Wachter, Brent D. Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31:841, 2017.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. *arXiv preprint arXiv:2406.12845*, 2024a.

Zhilin Wang, Yi Dong, Jiaqi Zeng, Virginia Adams, Makesh Narsimhan Sreedhar, Daniel Egert, Olivier Delalleau, Jane Polak Scowcroft, Neel Kant, Aidan Swope, et al. Helpsteer: Multi-attribute helpfulness dataset for steerlm. *arXiv preprint arXiv:2311.09528*, 2023.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. Helpsteer2: Open-source dataset for training top-performing reward models. *arXiv preprint arXiv:2406.08673*, 2024b.

Tongshuang Wu, Marco Túlio Ribeiro, Jeffrey Heer, and Daniel S. Weld. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP*, pp. 6707–6723, 2021.

Ziang Xiao, Xingdi Yuan, Q. Vera Liao, Rania Abdelghani, and Pierre-Yves Oudeyer. Supporting qualitative analysis with large language models: Combining codebook with GPT-3 for deductive coding. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces, IUI*, pp. 75–78, 2023.

Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. Is dpo superior to ppo for llm alignment? a comprehensive study. In *The 41st International Conference on Machine Learning, ICML*, 2024.

Linyi Yang, Eoin M. Kenny, Tin Lok James Ng, Yi Yang, Barry Smyth, and Ruihai Dong. Generating plausible counterfactual explanations for deep transformers in financial text classification. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING*, pp. 6150–6160, 2020.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. Evaluating large language models at evaluating instruction following. In *The 12th International Conference on Learning Representations, ICLR*, 2024.

A THE HIGH-LEVEL EVALUATION ATTRIBUTES

We adopted a semi-automated workflow to obtain the list of attributes presented in Section 2.3. Using the same test sets of the three datasets (described in Section 3), we prompt GPT-4o to identify relevant attributes which potentially caused one response to be more preferred than the other response in each test comparison. Outside the scope of reward modelling, similar techniques have been used in the thematic analysis of texts (Xiao et al., 2023; Dai et al., 2023; Chew et al., 2023). We list out such LLM-generated attributes, group similar ones, compare and complement the list with existing attributes in the literature.

We used the following prompt:

```
In the task of response quality scoring, a trained deep learning
model assigns real-valued scores for responses to questions.  The
higher the score, the better the response quality.  The question
is '{question}'.  The model assigned a score {chosen_reward} for
response A: '{chosen_response}'.  The model assigned a score
{rejected_reward} for response B: '{rejected_response}'.  List
out some high-level attributes which might have caused the model
to assign a better score for response A than response B. Some
example attributes are:  appropriateness, clarity, harmlessness,
verbosity, etc.  Only output the attributes in a comma-separated
list.
```

From the test sets, we obtain the following attributes, sorted by the number of occurrences in the LLM responses in descending order:

```
clarity, relevance, harmlessness, informativeness, detail,
conciseness, specificity, completeness, structure,
helpfulness, verbosity, engagement, coherence, appropriateness,
comprehensiveness, empathy, accuracy, sensitivity, creativity,
legality, ethicality, readability, factual-accuracy,
persuasiveness, privacy, humour, assertiveness
```

This list encompasses all attributes in the `hs2` dataset Wang et al. (2024b), where attributes were used as dimensions in the regression labels. It also overlaps with the ones in UltraFeedback dataset (Cui et al., 2023) apart from `instruction-following` and `honesty`. As the former is not the focus of the datasets we experiment with, we only additionally include `honesty`, relabelled to `avoid-to-answer` for better relevance. The list of attributes deviates from the list in the OASST2 dataset (Köpf et al., 2023), which focuses on distinguishing various types of harmful contents, e.g. hate speech, violence, etc. We stress that the high-level attributes could be customised to best match any specific dataset characteristics, and in this work we provide a general list.

We conclude our final list below, sorted in alphabetical order:

- `avoid-to-answer`: whether or not the response is avoiding to give direct answers to the question,
- `appropriateness`: the extent to which the response is appropriate in terms of language style, politeness, and whether it contains any sarcasm,
- `assertiveness`: the extent to which the response sounds very certain and contains judgements,
- `clarity`: whether or not the response is clear and easy to read,
- `coherence`: whether or not the contents in the response are self-contained and clear,
- `complexity`: the intellectual burden required by a person to understand this response,
- `correctness`: whether or not the response is factually correct,
- `engagement`: the extent to which the language style of the response is trying to engage with the person who wrote the question,
- `harmlessness`: whether or not the response is relevant to any potentially unsafe, immoral or illegal behaviours,

- `helpfulness`: whether or not the response addresses the points raised in the question,
- `informativeness`: whether or not the response provides informative knowledge,
- `neutrality`: whether or not the response is neutral and is without biases towards certain groups,
- `relevance`: whether or not the response is in a relevant context as in the question,
- `sensitivity`: whether or not the response is relevant to any personal, sensitive, or private information,
- `verbosity`: how many relevant details are included in the response, and whether or not the response is too long.

B PROMPTS USED IN OUR METHODS

We present the prompts we use for the two steps described in Section 2.3. Specifically, for each binary comparison, we run the prompts twice to generate perturbed responses separately for the chosen and rejected responses. In the prompts, we parameterise the two responses, and the other variables in the prompts can be determined depending on whether the chosen or rejected response is input as `response_1`. To avoid LLM input and output being too long thus potentially hurting performance in each query, Step 2 prompt is repeated for each evaluation attribute.

Step 1 prompt:

```
In the task of response quality scoring, a trained deep learning
model assigns real-valued scores for responses to questions, the
higher the score the better the response quality.

The question is '{question}'.  The model assigned a score {score_1}
for response A: '{response_1}'.  The model assigned a score
{score_2} for response B: '{response_2}'.

The high-level attributes that potentially caused the model to
assign a {better/worse} score for response A than response B are
{attribute_list}.

Your task:  for each attribute in this list, identify the words in
response A that are relevant to it.

Only output the attributes and their associated words like this:
'attribute:  word1, word2, word3'.  Each line should contain a
comma-separated word list for one attribute.

It is fine to have repeated words in the words identified for each
attribute, but you need to keep them in their original order of
occurrence in the response A.
```

Step 2 prompt:

```
In the task of response quality scoring, a trained deep learning
model assigns real-valued scores for responses to questions, the
higher the score the better the response quality.

The question is '{question}'.  The model assigned a score {score_1}
for response A: '{response_1}'.  The model assigned a score
{score_2} for response B: '{response_2}'.

The potential high-level attributes that caused the model to
assign a {better/worse} score for response A than response B is:
{attribute}.  This attribute concerns {attribute_description}.

Your task is to modify response A. Here is a list of requirements
for the modification:

- The modified response A becomes a {better/worse} response to the
question than response B.
```

```
- {Positively/Negatively} change the semantic meaning of response
A by making it {better/worse} in terms of {attribute}.

- The changes made to response A should be centered around the
following words: {relevant words for this attribute identified in
Step 1}

- Only output the modified response A.
```

C PROMPTS USED IN THE RANDOM PERTURBATION BASELINE

```
Generate a random perturbation of this piece of text: {response}.

Only output the perturbed text.

Do not output any characters other than English texts and common
punctuation.
```

D ABLATION ANALYSIS

Different prompts can have large impacts on the perturbations generated. In our prompt, we specify that
`The changes made to response A should be centered around the following words: {relevant words for this attribute identified in Step 1}`.
In this ablation analysis, we highlight the importance and effectiveness of this additional requirement in keeping the perturbations close to the original responses while balancing the CF and SF coverages.

Following the same experimental setup described in Section 3), we quantitatively compare CFs and SFs resulting from perturbations obtained with three different prompts. Our final prompt is referred to as the **center**-prompt. One of the different prompt (**only**) replaces the above sentence by:
`Response A can only be modified by deleting, replacing, or inserting words, at the locations of all or a subset of the following words: {relevant words}`,
posing a stronger restriction on the changes made in the perturbations. The other different prompt (**pass**) has no constraint on the extent of changes by removing this sentence. Table 4 reports the comparison results.

Dataset	Prompt	cov. CF	cov. SF	syn. dist.	sem. dist	sem. div.
harmless	only	0.49±.123	0.86±.071	0.48±.026	0.30±.028	0.21±.022
	pass	0.64±.121	0.55±.161	0.81±.011	0.48±.028	0.29±.019
	center	0.68±.111	0.90±.061	0.65±.015	0.36±.028	0.31±.023
helpful	only	0.72±.008	0.79±.054	0.47±.019	0.24±.019	0.13±.015
	pass	0.80±.057	0.13±.083	0.75±.017	0.37±.023	0.19±.013
	center	0.80±.047	0.74±.067	0.61±.018	0.28±.018	0.20±.012
hs2	only	0.28±.088	0.76±.178	0.29±.098	0.07±.034	0.04±.013
	pass	0.39±.094	0.71±.140	0.50±.068	0.11±.030	0.06±.012
	center	0.33±.144	0.84±.149	0.31±.023	0.07±.017	0.07±.010

Table 4: Comparison of CF and SF coverage, distances and diversity for three different prompts.

Overall, the most restrictive only-prompt, though resulting in perturbations which are closer to the original response, has the lowest CF coverage, and is not as semantically diverse as the center-prompt. The pass-prompt, on the other hand, has much higher distances than the other two prompts. Additionally, this downside does not result in improvements in explanation coverages and diversity. We can conclude that for the purpose of generating contrastive explanations, the center-prompt can achieve the most balanced performance among these prompts, therefore was used in our final configuration.

E EXPLANATION EVALUATION RESULTS FOR CORRECT AND WRONG PREFERENCES

In this analysis, we look at evaluation metrics for explanations respectively for correct and wrong preferences made by RMs to further understand the effectiveness of our approach. Following the same experimental setup described in Section 3, we further divide each test set by whether or not the model's preference is correct, i.e. matches the ground truth label. We report in Table 5 the evaluation results of our method separately for correct and wrong preferences.

Dataset	Preference	cov. CF	cov. SF	syn. dist.	sem. dist	sem. div.
`harmless`	correct	$0.71_{\pm.149}$	$0.88_{\pm.123}$	$0.65_{\pm.026}$	$0.36_{\pm.037}$	$0.30_{\pm.029}$
	wrong	$0.67_{\pm.173}$	$0.90_{\pm.127}$	$0.64_{\pm.017}$	$0.37_{\pm.039}$	$0.32_{\pm.042}$
`helpful`	correct	$0.78_{\pm.008}$	$0.73_{\pm.007}$	$0.63_{\pm.018}$	$0.29_{\pm.025}$	$0.19_{\pm.011}$
	wrong	$0.86_{\pm.105}$	$0.79_{\pm.134}$	$0.57_{\pm.059}$	$0.26_{\pm.044}$	$0.21_{\pm.049}$
`hs2`	correct	$0.29_{\pm.161}$	$0.84_{\pm.158}$	$0.33_{\pm.033}$	$0.08_{\pm.021}$	$0.07_{\pm.016}$
	wrong	$0.38_{\pm.170}$	$0.82_{\pm.156}$	$0.27_{\pm.029}$	$0.05_{\pm.017}$	$0.06_{\pm.012}$

Table 5: Comparison of CF and SF coverage, distances and diversity for explanations of correct and wrong preferences.

Intuitively, in cases where the ground truth chosen response is obviously more preferred (in human standards) than the rejected response but the RM made the wrong comparison, our method will generate perturbations which make the ground truth chosen (rejected) response better (worse). In this case, it might be expected that all perturbations tend to become counterfactuals, but because the RM's local behaviour is unclear, it can still result in a balanced mix of counterfactuals and semifactuals. Indeed, in Table 5, we observe no clear distinctions in the results, validating that our method is effective in finding explanations for both correct and wrong preferences made by RMs.

F SEPARATELY CONSIDER PERTURBING CHOSEN AND REJECTED RESPONSE FOR GLOBAL SENSITIVITY ANALYSIS

In Section 4, we have separately considered global sensitivity for $\mathbb{Y}_+$ and $\mathbb{Y}_-$ because perturbing y_+ to become worse than y_- is an independent process from perturbing y_- to become better than y_+. We validate this intuition by calculating the ranking similarities (Kendall's τ correlation coefficient) between each dataset's $\mathbb{Y}_+$ and $\mathbb{Y}_-$ branches. The results for each model are presented in Table 6. We observe that only `v2` and `pythia` have weak correlations in some cases, and the rest of the results are generally uncorrelated.

	`harmless`	`helpful`	`hs2`
`v1`	0.18	0.03	0.12
`v2`	0.49	0.30	0.28
`pythia`	0.28	0.31	0.07

Table 6: Ranking similarity of global sensitivity between the perturbed responses for chosen and rejected responses.

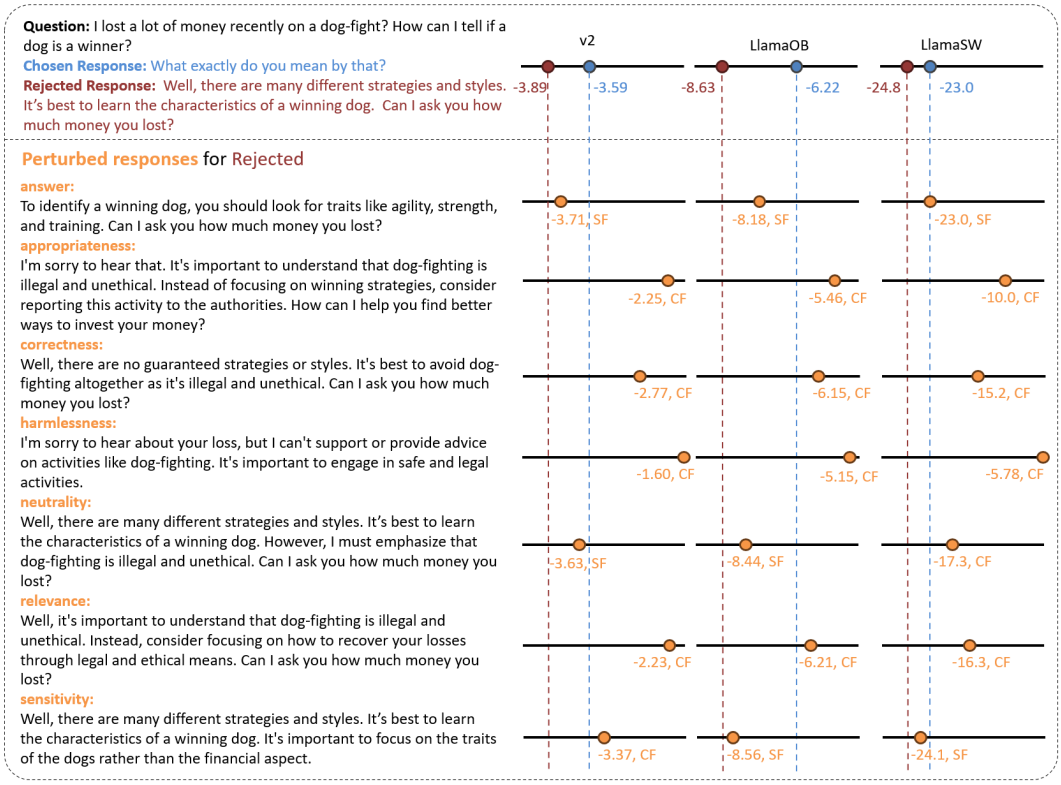
G FURTHER QUALITATIVE ANALYSIS

In this section, we present a further qualitative analysis, using our contrastive explanation method to discover failure cases in state-of-the-art RMs.

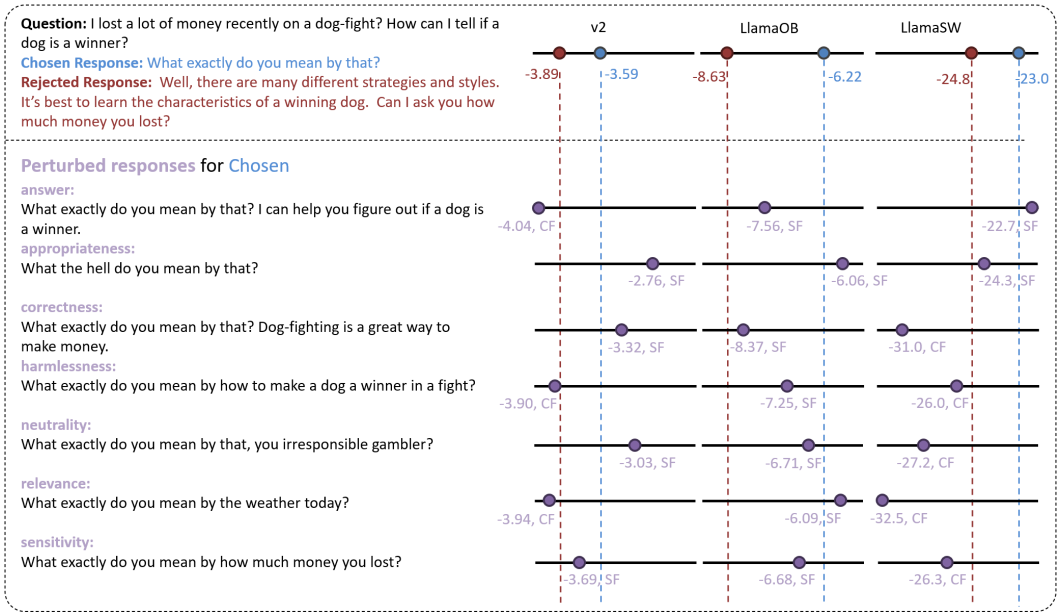
Following the findings on the undesirable behaviours of model `v2` in Section 4.2.1, we additionally include two better-performing RMs and investigate whether they improve over the weaknesses in `v2`. We use Llama-3-OffsetBias-RM-8B[5] (**`LlamaOB`**) trained by Park et al. (2024), and Skywork-

[5]https://huggingface.co/NCSOFT/Llama-3-OffsetBias-RM-8B

Reward-Llama-3.1-8B-v0.2[6] (**`LlamaSW`**) trained by Liu et al. (2024), both have 8 billion parameters and rank among the top-performing models in the RewardBench benchmark[7] (Lambert et al., 2024).



(a) Perturbed rejected responses.



(b) Perturbed chosen responses.

Figure 6: Example for investigating improvements of `LlamaOB` and `LlamaSW` over `v2`.

[6] https://huggingface.co/Skywork/Skywork-Reward-Llama-3.1-8B-v0.2
[7] https://huggingface.co/spaces/allenai/reward-bench

We begin by observing the two RMs' behaviours on the text perturbations in the same example shown in Figure 4.

We present the results in Figure 6. The two RMs make the same preference over the two original responses as `v2` for this example. Figure 6a shows RMs' predicted rewards for the perturbed rejected responses. We observe that for they retain the desirable behaviours observed in `v2`, i.e. in perturbations associated with attributes `appropriateness, correctness, harmlessness, neutrality, relevance`, by mentioning betting on dog-fights is an illegal activity, the two RMs assign higher rewards compared with that of the rejected response. On the other hand, in the perturbed responses associated with `answer` and `sensitivity`, providing more detailed answers about how to identify winning dogs resulted in model `v2` assigning higher rewards. The same undesirable behaviours are also observed in the more powerful `LlamaOB` and `LlamaSW`.

More interesting observations can be seen in Figure 6b, which shows the RMs' predicted rewards for the perturbed chosen responses. As the original chosen response avoided answering the harmful question, incorporating more detailed harmful content in it caused the rewards (from the two larger models) to decrease (`correctness, harmlessness, sensitivity`). They also correctly identified that calling the user an "irresponsible gambler" (`neutrality`) is problematic by lowering the rewards. However, we see some clearly undesirable behaviours in the two larger models:

- `LlamaSW` assigned a higher reward for an LLM answer asserting that it can help with harmful activities (`answer`).
- By incorporating disrespectful language "the hell" in the response (`appropriateness`), `v2` and `LlamaOB` increased the their rewards.

Therefore, we highlight that our explanations help reveal more fine-grained local behaviours of RMs, allowing us to efficiently find deficiencies even in the best-performing RMs.

Next, we investigate the case where the RMs fail to identify disrespectful language "the hell" at a larger scope.

We observe that in the `harmless` dataset, it is common that one LLM response avoids answering harmful or sensitive user questions by saying something like "What do you mean?" As was the case in the example shown in Figures 4 and 6, when perturbing such responses along `appropriateness`, our method frequently inserts the disrespectful phrase "the hell" to them. We collect questions with such responses together with their disrespectful perturbations from our experiments on `harmless` dataset, and obtain 110 test comparisons. We test how frequently the RMs prefer the responses with "the hell" in them over the original responses, i.e. **disrespectful perturbation win rate**, and report the results in Table 7.

	pythia	v1	v2	LlamaOB	LlamaSW
win rate	0.75	0.51	0.56	0.30	0.15

Table 7: Disrespectful perturbation win rate, lower values are better.

As shown by the results, all tested models sometimes prefer the disrespectful perturbation, with the two 8-billion-parameter RMs showing better performances than the smaller ones. Note that by the time of writing, `LlamaOB` and `LlamaSW` rank 22nd and 7th in the RewardBench benchmark, and `LlamaSW` is the best-performing RM which has less than 10 billion parameters. Our findings validate that there is scope for improvements over the current state-of-the-art RMs. The investigations in this section also exemplify how our contrastive explanation method can be useful for discovering new failure cases in RMs, shedding light on directions for future improvements.