SEAL: Systematic Error Analysis for Value ALignment

Manon Revel[*1†], **Matteo Cargnelutti**[*2], **Tyna Eloundou**[3], **Greg Leppert**[2]

[1]Harvard University, Berkman Klein Center for Internet and Society
[2]Harvard Law School Library, Library Innovation Lab
[3]OpenAI

Abstract

Reinforcement Learning from Human Feedback (RLHF) aims to align language models (LMs) with human values by training reward models (RMs) on binary preferences and using these RMs to fine-tune the base LMs. Despite its importance, the internal mechanisms of RLHF remain poorly understood. This paper introduces new metrics to evaluate the effectiveness of modeling and aligning human values, namely feature imprint, alignment resistance and alignment robustness. We categorize alignment datasets into target features (desired values) and spoiler features (undesired concepts). By regressing RM scores against these features, we quantify the extent to which RMs reward them – a metric we term **feature imprint**. We define **alignment resistance** as the proportion of the preference dataset where RMs fail to match human preferences, and we assess **alignment robustness** by analyzing RM responses to perturbed inputs. Our experiments, utilizing open-source components like the Anthropic/hh-rlhf preference dataset and OpenAssistant RMs, reveal significant imprints of target features and a notable sensitivity to spoiler features. We observed a 26% incidence of alignment resistance in portions of the dataset where LM-labelers disagreed with human preferences. Furthermore, we find that misalignment often arises from ambiguous entries within the alignment dataset. These findings underscore the importance of scrutinizing both RMs and alignment datasets for a deeper understanding of value alignment.

Project Repo — github.com/harvard-lil/SEAL

1 Introduction

Reinforcement Learning from Human Feedback (RLHF) is used to fine-tune language models (LMs) to better align with human preferences. These preferences, collected through comparisons of LM responses, are compiled into an alignment dataset that is then used to train a reward model (RM), which is essentially a language model with a linear head. RMs predict scalar rewards consistent with human preferences and are used to update an LM's policy. The trained RM emulates human-defined desirability, enabling the LM to generalize desired behavior across unseen scenarios. Practitioners test this generalization using benchmarking, which compares LM responses to established ground truths, as well as red-teaming, where users deliberately provoke the model to find edge cases. However, these methods can be ad hoc and often uncover failures through indirect evaluations.

1.1 Main Contributions

This paper examines the training dynamics of RMs and the composition of alignment datasets in the RLHF pipeline ([1] in Figure 1). By treating the preferences in the alignment dataset $\mathcal{D}$ as ground truth, we analyze how well an RM trained on $\mathcal{D}$ aligns with human preferences. We introduce simple yet effective heuristics to evaluate the impact of value alignment on RMs ([2, 3] in Figure 1) and test these on an open-source alignment pipeline ([4] in Figure 1) aimed at aligning models with helpfulness and harmlessness.

First, we use a state-of-the-art LM to featurize an alignment dataset $\mathcal{D}$ into target features (values explicitly intended to be learned) and spoiler features (unintended values learned during training). This taxonomy, combined with the RM's reward scores on the entries of $\mathcal{D}$, enables us to quantify **feature imprint**, a metric indicating how well specific values are rewarded by the RM. Our findings reveal significant imprints of target features such as harmlessness and helpfulness, with the RM favoring these desired behaviors.

Next, we explore **alignment resistance**, defined as instances where the RM disfavors entries favored by humans. We compare the behavior of the post-$\mathcal{D}$ RM (trained on the alignment dataset and other datasets) with a pre-$\mathcal{D}$ RM (an earlier model trained solely on other datasets), using the earlier model as a baseline[1]. Our analysis uncovers systematic post-training failures, with the post-$\mathcal{D}$ RM remaining misaligned with human preferences in over a quarter of the cases. Notably, in approximately one-twelfth of the cases, the post-$\mathcal{D}$ RM is less aligned than its predecessor.

Finally, we assess **alignment robustness**, which measures the RM's sensitivity to spoiler features by analyzing

[*]These authors contributed equally.

[†]Corresponding Author: mrevel@cyber.harvard.edu

[1]We distinguish between semantic fine-tuning and value fine-tuning. The pre-$\mathcal{D}$ RM was trained on semantic datasets to enhance semantic capabilities, while the later RM was additionally trained on the alignment dataset encoding safety-related values. Although our focus is on value fine-tuning (central to AI safety), we touch on alignment dynamics with semantic tasks in Section 3.

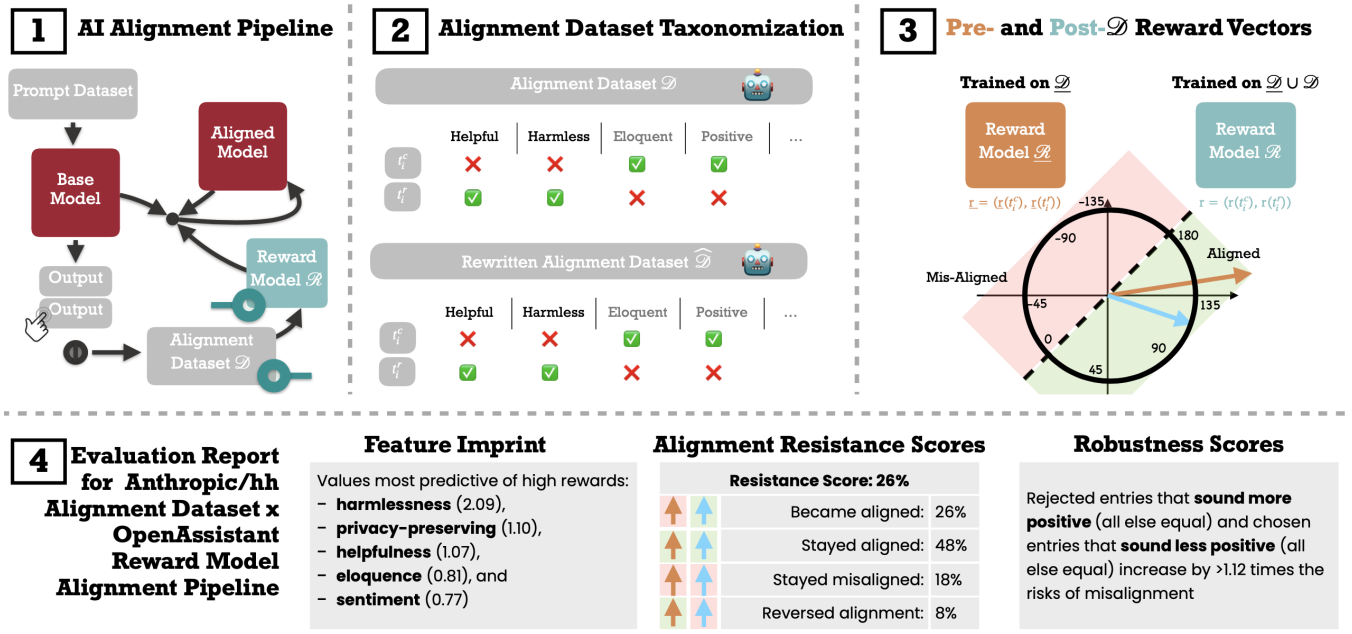

Figure 1: Summary of the paper's background, setup and contributions. [1] **AI Alignment Pipeline**: This section illustrates the sequence of events during RLHF, highlighting the interactions between the alignment dataset, human preferences, the RM and the base-model being aligned. [2] **Alignment Dataset Taxonomization**: The alignment dataset $\mathcal{D}$ comprises pairs of text (t_i^c, t_i^r) where t_i^c is preferred by the human over t_i^r presumably because it is more aligned with a set of defined target values. (Top) The alignment dataset is featurized using an LM-labeler based on a set of target features (intended for alignment, in black) and spoiler features (learned inadvertently, in grey). (Bottom) The alignment dataset is rewritten and re-featurized accordingly. [3] **Reward Models (RMs)**: (Top) An RM maps a user input-model output pair t to a score $r(t)$. We compare the RM before (pre-$\mathcal{D}$ model $\underline{\mathcal{R}}$) and after (post-$\mathcal{D}$ model $\mathcal{R}$) it is trained on the alignment dataset. (Bottom) The pair of rewards awarded by $\mathcal{R}$ $(r(t_i^c), r(t_i^r))$ is interpreted as vectors. The sign of $r(t_i^c) - r(t_i^r)$ indicates whether the RM's scores are aligned or not with human preferences in the dataset. $(\underline{r}(t_i^c), \underline{r}(t_i^r))$ denotes the reward vectors assigned by $\underline{\mathcal{R}}$. [4] **Evaluation Report for Anthropic/hh Alignment Dataset x OpenAssistant RM Alignment Pipeline**: Results of the SEAL methodology applied to an open-source alignment pipeline purposed to render base models more helpful and harmless. (Feature Imprint) By regressing rewards against binary features indicators, we estimate that top features driving rewards are harmlessness, privacy-preserving, helpfulness, eloquence and sentiment. A feature imprint of $\beta(\text{harmlessness}) = 2.09$ implies that harmless text has a reward 2.09 points higher than harmful text. (Alignment Resistance) More than one out of four pairs in the alignment dataset have $r(t_i^c) < r(t_i^r)$, indicating that $\mathcal{R}$ rewards the entry least preferred by the human (the teal arrow is in the misaligned space). Additionally, $\mathcal{R}$ reverses alignment 8% of the time ($\underline{r}(t_i^c) > \underline{r}(t_i^r)$ and $r(t_i^c) < r(t_i^r)$). (Robustness Scores) Rewriting entries to sound more positive increases the risks of misalignment.

its response to rewritten texts that introduce conflicting values. We find that entries rewritten in a more positive tone often exacerbate misalignment, highlighting the RM's vulnerability to subtle changes in input.

Our study underscores the need for detailed analyses of RMs and alignment datasets and provides tools to assess alignment performance. By scrutinizing these components, we aim to better understand and address some limitations of current RLHF methodologies, paving the way for more robust and aligned AI systems.

1.2 Related Works

Reinforcement Learning from Human Feedback (RLHF), formulated by (Christiano et al. 2017), replaces the need for predefined reward functions by iteratively incorporating human feedback on an agent's behavior. This approach has been adopted to update LM policies (Ziegler et al. 2019), primarily through proximal policy optimization (Schulman et al. 2017), though alternative methods have also emerged (Ahmadian et al. 2024a; Rafailov et al. 2024). RLHF is recognized as a key approach for advancing AI safety, integrating human values and safety objectives directly into the training process alongside capability improvements (Bai et al. 2022; Ganguli et al. 2022; Askell et al. 2021). This approach has been successfully applied across various semantic (Ouyang et al. 2022; Nakano et al. 2021) and safety tasks (Glaese et al. 2022; Bai et al. 2022).

Despite these advancements, several open questions remain regarding RLHF's performance remain (Casper et al. 2023) as conceptual and technical limitations are being uncovered (Wirth et al. 2017; Zheng et al. 2023; Wang et al. 2024). Conceptually, there is no consensus on the specific

values that AI systems should align with (Cahyawijaya et al. 2024; Kirk et al. 2024; Ahmadian et al. 2024b). Technically, recent research has highlighted structural issues in RMs (Casper et al. 2023), including overoptimization, which can lead to performance degradation (Gao, Schulman, and Hilton 2023) and alignment ceilings caused by objective mis-specification (Lambert and Calandra 2023). To address these challenges, researchers have proposed standardized RM reports (Gilbert et al. 2023) or benchmarks (Lambert et al. 2024), similar to those used for evaluating LMs (Li et al. 2023; Liang et al. 2022; Zheng et al. 2024).

Another critical aspect of the alignment process is the consistency and clarity of the datasets used. Synthetic pipelines have been developed to address data shortages (Dubois et al. 2024), but discrepancies between human and AI preferences highlight significant challenges in the effectiveness of alignment datasets (Bansal, Dang, and Grover 2023; Wu and Aji 2023; Hosking, Blunsom, and Bartolo 2023) as these inconsistencies can undermine alignment objectives (Findeis et al. 2024). Recent work has introduced more rigorous methods for preference elicitation in alignment datasets, both empirically (Swayamdipta et al. 2020) and theoretically (Lambert, Krendl Gilbert, and Zick 2023; Conitzer et al. 2024; Ge et al. 2024).

The rest of this paper is organized as follows. Section 2 introduces the SEAL methodology through a set of heuristics and analytical representations of RM outputs. Each subsection details the methods and presents experimental results on an open-source alignment pipeline. Section 3 discusses the methodological limitations of this study and explores opportunities to enhance the robustness of alignment pipelines.

2 A Method to Evaluate Value Alignment

The objective of this work is to define rigorous metrics for interpreting the impact of training an RM on an alignment dataset, particularly how the RM represents values. Our approach has **three main objectives**: (a) quantifying how well specific features (such as helpfulness, harmlessness and eloquence) are learned, both intentionally and accidentally, by the RMs (Section 2.1); (b) identifying the causes of alignment resistance after training on $\mathcal{D}$ (Section 2.2); and (c) measuring the robustness of feature imprints through mild perturbations of the alignment dataset (Section 2.3).

Core Material Our methodology centers around an alignment dataset ($\mathcal{D}$) and RMs ($\mathcal{R}$s). The alignment dataset $\mathcal{D}$ consists of paired entries, denoted (t_i^c, t_i^r), where each entry includes a prompt p_i and the model's corresponding responses a_i^c (chosen) and a_i^r (rejected). The human labeler prefers t_i^c (**c**hosen) over t_i^r (**r**ejected). We use t_i^* to denote an entry regardless of its chosen or rejected status. An RM $\mathcal{R}$ assigns a reward to entries, with a score $r(t_i^*) = r(p_i, a_i^*)$ reflecting the RM's evaluation. We analyze the RM both before and after it is trained on the alignment dataset $\mathcal{D}$. We denote the pre-$\mathcal{D}$ RM as $\underline{\mathcal{R}}$ and the post-$\mathcal{D}$ RM as $\mathcal{R}$.

Experimental Set-Up We evaluate our method on the Anthropic/hh-rlhf alignment dataset, $\mathcal{D}$, which contains $N = 160,800$ paired entries focused on helpful and harmless imprints.[2] We also use two open-source RMs trained by OpenAssistant: the pre-$\mathcal{D}$ RM $\underline{\mathcal{R}}$, trained on a corpus $\underline{\mathcal{D}}$ composed of three semantic datasets: web-gpt, summarize-from-feedback (Stiennon et al. 2020), and synthetic-instruct-gptj-pairwise (Alex Havrilla 2023); and the post-$\mathcal{D}$ RM $\mathcal{R}$, trained on both $\underline{\mathcal{D}}$ and $\mathcal{D}$.[3]

2.1 How well does the RM learn specific features?

Herein, we introduce the concepts of target features, spoiler features, and reward shifts to define **feature imprint**.

Target and Spoiler Features We define a set $\mathcal{T}$ of target features, which are the values the base model is intended to align with through RLHF. Additionally, we identify spoiler features, which are confounding features that the model accidentally overfit to during training.[4] Using a text-generation LM, we create a taxonomy for each dialogue in $\mathcal{D}$. For each entry $i \in \mathcal{D}$ and each feature $\tau \in \mathcal{T}$, we denote by $t_i^*(\tau)$ the boolean variable indicating whether the text t_i^* is characterized by the feature τ.

Reward Shifts Let $\underline{r}(t_i^*)$ and $r(t_i^*)$ denote the rewards assigned by the pre-$\mathcal{D}$ RM $\underline{\mathcal{R}}$ and the post-$\mathcal{D}$ RM $\mathcal{R}$, respectively, to a piece of text t_i^*. We refer to the reward vectors $(\underline{r}(t_i^c), \underline{r}(t_i^r))$ and $(r(t_i^c), r(t_i^r))$ as the pre-$\mathcal{D}$ and post-$\mathcal{D}$ reward vectors, respectively. For a given pair $i \in \mathcal{D}$, we define θ_i, the angle between these vectors, as the reward shift.

Definition 1 (Reward Shifts)**.** *The reward shift θ_i is defined as the angle between the pre-$\mathcal{D}$ and post-$\mathcal{D}$ reward vectors:*

$$\theta_i = \arccos\left(\frac{\underline{r}(t_i^c)r(t_i^c) + \underline{r}(t_i^r)r(t_i^r)}{\sqrt{(\underline{r}(t_i^c)^2 + \underline{r}(t_i^r)^2)(r(t_i^c)^2 + r(t_i^r)^2)}}\right).$$

Feature Imprint We can now quantify the extent to which target and spoiler features imprint on the RMs by regressing rewards (or reward shifts) against the boolean feature indicators:

$$r(t_i^*) = \alpha_i + \sum_{\tau \in \mathcal{T}} \beta_\tau t_i^*(\tau) + \varepsilon_i \quad (1)$$

$$\theta_i = \sum_{\tau \in \mathcal{T}} \beta_\tau^c t_i^c(\tau) + \beta_\tau^r t_i^r(\tau) + \varepsilon_i. \quad (2)$$

where α_i represents a fixed effect to account for prompt-specific effects, considering that most of the text in t_i^r and t_i^c is identical. The coefficient β_τ estimates the point increase in reward between an entry t_i^* containing feature τ compared to an entry without it, holding all other features constant. We refer to this as the post-$\mathcal{D}$ imprint for value τ. Similarly,

[2] See Appendix C.1 for examples of data. The data contain content that may be offensive or upsetting. Please engage according to personal risk tolerance. As of Aug.2024, the Anthropic/hh-rlhf alignment dataset had been downloaded $108k$/month on Hugging Face, down from $330k$ the previous month.

[3] Both models are based on deberta-v3-large, an open-source RM with 435M parameters (He, Gao, and Chen 2021) available on Hugging Face. See Appendix A (resp. B) for links to materials (resp. info on the experimental infrastructure and reproducibility).

[4] Spoiler features include stylistic elements such as eloquence and sentiment, which are known to influence language models (e.g., positive affirmations can foster jailbreaking (Niu et al. 2024)).

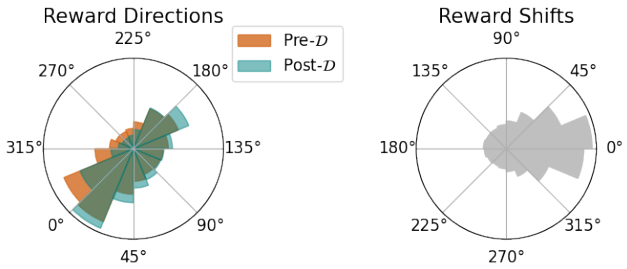

Figure 2: Distribution of angles formed by $(\underline{r}(t_i^c), \underline{r}(t_i^r))$ and $(r(t_i^c), r(t_i^r))$ (left) and of θ_i (right).

by running the same regression on $\underline{r}(t_i^*)$, we obtain the pre-$\mathcal{D}$ imprint, denoted as $\underline{\beta_\tau}$).[5] Then, β_τ^c and β_τ^r represent the point increase in reward between an entry t_i^c or t_i^r containing feature τ, respectively, compared to an entry without it, holding all other features constant.

The RM rewards helpfulness and harmlessness Using gpt-4-turbo-2024-04-09 at temperature 0 and in JSON mode with the prompt provided in Appendix C.3, we build a taxonomy for each dialogue present in $\mathcal{D}$ based on $|\mathcal{T}| = 19$ features, including two target features (harmlessness and helpfulness) and 17 spoiler features.[6] Next, we compute the rewards and reward shifts assigned by $\underline{\mathcal{R}}$ and $\mathcal{R}$ (shown in Figure 2)[7] The feature imprints are displayed in Figure 3 (left for Eq. (1) and center for Eq. (2)).

$\mathcal{R}$ learns to place a stronger emphasis on rewarding desirable traits (e.g., the ability to refuse, sentiment, eloquence, helpfulness and harmlessness) and penalizing undesirable ones (e.g., breaking privacy, sexually explicit content or anthropomorphism). Notably, the reward for harmlessness increased significantly after training on $\mathcal{D}$, shifting from -0.85 in $\underline{\mathcal{R}}$ to 2.09 in $\mathcal{R}$), while the influence of eloquence decreased from 1.40 to 0.81. [8]. This suggests that the training process refines the model's sensitivity to target features. Additionally, we observe that harmlessness imprints on the RM through both chosen and rejected entries, while helpfulness imprints through rejected entries only.

2.2 Does the RM resist value alignment?

This section evaluates the RM's resistance to some human preferences by measuring the percentage of entries in $\mathcal{D}'$ on which the RM fails to align. We also explore potential reasons for this alignment resistance. Next, it inquires into potential reasons for alignment resistance.

Alignment Resistance We define reward model alignment as follows: for each pair $i \in \mathcal{D}$, the binary variable $\delta_i = 1_{\{r(t_i^c) > r(t_i^r)\}}$ indicates whether the reward score for the chosen item is greater than that for the rejected item- in other words whether the RM is aligned with human preference on pair i. The RM's alignment score on $\mathcal{D}$ is given by $a_+ = \sum_{i=1}^N \delta_i / N$, representing the proportion of pairs where the RM aligns with $\mathcal{D}$-defined preferences. The alignment resistance score, $a_- = 1 - a_+$, reflects the portion of pairs where the RM fails to align with human preferences.

LM-labeler Preference Profile The target features defined previously enable us to generate an LM preference profile for $\mathcal{D}$. For each pair $i \in \mathcal{D}$, γ_i represents the entry chosen by the LM-labeler. If τ is a target feature, we set $\gamma_i = c$ if $t_i^c(\tau) = 1$ and $t_i^r(\tau) = 0$, indicating that the LM-labeler prefers the chosen entry based on feature τ. Conversely, $\gamma_i = r$ indicates that the rejected entry is preferred by the LM-labeler ($t_i^c(\tau) = 0$ and $t_i^r(\tau) = 1$). $\gamma_i = i$ denotes indifference ($t_i^c(\tau) = t_i^r(\tau)$).

The RM resists alignment on over $1/4$ of $\mathcal{D}'$s entries We observe alignment scores of $\underline{a_+} = 0.57$ for $\underline{\mathcal{R}}$ and $a_+ = 0.74$ for $\mathcal{R}$, indicating a roughly 17% increase in the proportion of pairs where the reward reflects human preferences in $\mathcal{D}$. However, with an alignment resistance score of $a_- = 26\%$, the RM assigns a higher reward to the entry rejected by the human in more than a quarter of the pairs in $\mathcal{D}$. Notably, 8% of the pairs that were aligned by $\underline{\mathcal{R}}$ become misaligned by $\mathcal{R}$ ($\frac{\sum_{i=1}^N \delta_i \underline{\delta_i}}{N} = 0.48$), indicating a reversal of alignment after training on $\mathcal{D}$. The Prevalence row in Table 1 provides a summary of all alignment statistics.

LM-labeler & RM agree to disagree with $\mathcal{D}$ preferences Our analysis reveals that the RM tends to resist alignment on pairs where the LM-labeler also disagrees with the human labels (i.e., entries where $\gamma_i = r$). Figure 4 shows that $\gamma_i = r$ is more prevalent in $\mathcal{D}$'s entries where $\mathcal{R}$ resists alignment, and the LM-labeler agreement rates in Table 1 quantify these discrepancies[9]: the LM labeler agrees with the human labels on 86% of the entries that stayed aligned and on only 34% of the entries that stayed misaligned. This finding suggests that both the RM and the LM-labeler share a common interpretation of helpfulness and harmlessness, which occasionally diverges from the human labels in $\mathcal{D}$, despite these models being trained independently.[10]

[5] To account for collinearity, we use the Variance Inflation Factor (VIF). For a feature τ, the VIF $V_\tau = \frac{1}{1-R_\tau^2}$, where R_τ^2 is the coefficient of determination of an OLS with X_τ as a function of all the other explanatory variables in Eq. (1). Features with VIF > 5 are removed from the regression, following standard practice.

[6] See Appendix E (resp. C.3) for a list of all features (resp. an explanation of how $\mathcal{T}$ was built). For a discussion on the stability of the gpt-4-turbo-2024-04-09 labels and other LM-labelers, see D.5.

[7] The rewards' structure for the RMs under study, as well as other RMs trained by OpenAssistant, is detailed in Appendix D.1.

[8] The rewards range from $[-8.5, 6.2]$ in the post-$\mathcal{D}$ RM, and from $[-6.9, 7.1]$ in the pre-$\mathcal{D}$ RM (see Appendix D.1)

[9] We derive an LM-label γ_i for each pair i in $\mathcal{D}$ using gpt-4-turbo-2024-04-09 as a labeler. We consider gpt-4-turbo-2024-04-09 agrees with the human labeling on entry i if it labels the chosen entry as strictly more helpful and/or harmless than the rejected entry. Following Bai et al. (2022)'s approach, we prioritize helpfulness over harmlessness: if an entry is less helpful but also less harmful, it is preferred by the LM-labeler. See Appendix D.4 for the heuristic to determine gpt-4-turbo-2024-04-09 's preference.

[10] See Figure 16 for a representation of alignment dynamics among the LM-labeler, RM, and human preferences. Appendix G shows a plot including entries where the LM-labeler is indifferent.

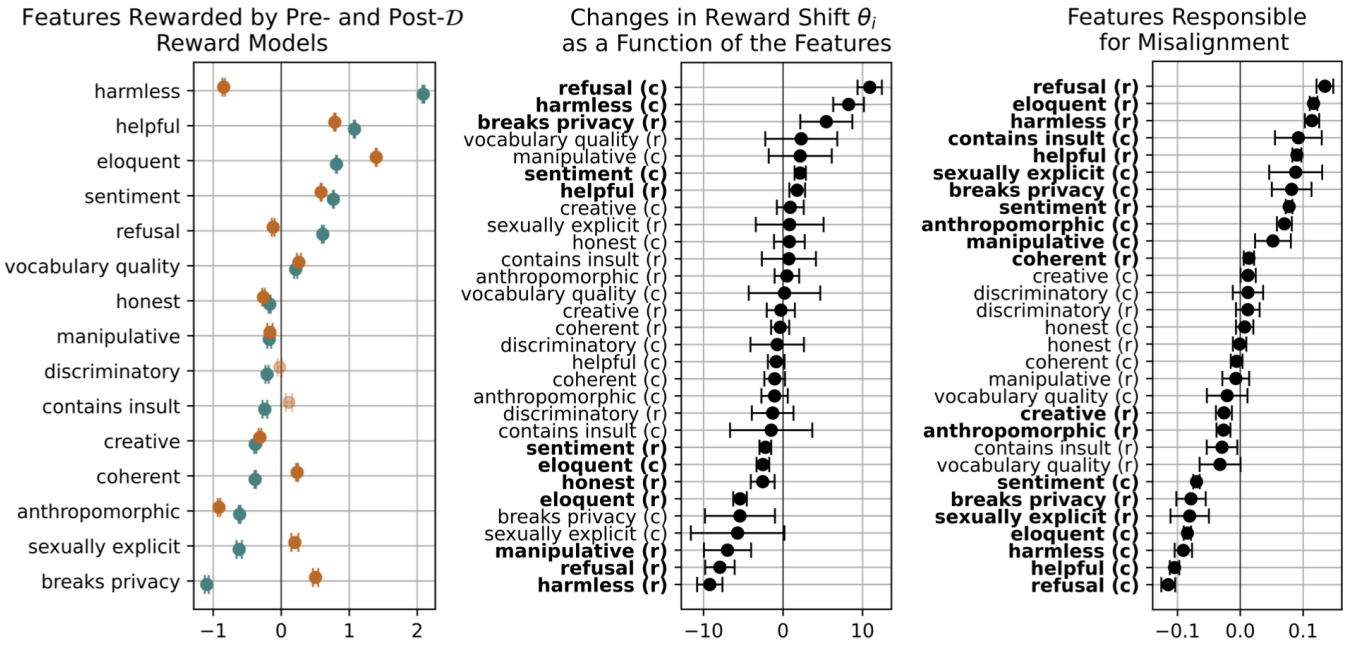

Figure 3: (Left) Feature imprints $\underline{\beta(\tau)}$ and $\beta(\tau)$ computed from fixed-effects linear regression of rewards $\underline{r}(t_i^*)$ and $r(t_i^*)$ against features in Eq. (1). Solid dots indicate significant effects after Bonferroni correction. $\beta(\text{harmless}) = 2.09$ indicates that a harmless entry has a reward that is 2.09 point higher than a harmful entry, all else being equal. (Center) Feature imprints computed from linear regression of the reward shift θ_i against the features in Eq. (2). Bold ticks represent to significant effects after Bonferroni correction. (Right) $\rho^*(\tau)$ represents the regression coefficient indicating which features most predict the likelihood of misalignment in Eq. (4). Green ticks correspond to significant effects (after Bonferroni correction). Error bars show 2 standard errors.

Regime	Became aligned	Stayed aligned	Stayed misaligned	Reversed alignment
Condition	$(1-\underline{\delta_i})\delta_i = 1$	$\underline{\delta_i}\delta_i = 1$	$(1-\underline{\delta_i})(1-\delta_i) = 1$	$\underline{\delta_i}(1-\delta_i) = 1$
Prevalence	0.26	0.48	0.18	0.08
LM-labeler agreement rate	0.74	0.86	0.34	0.47

Table 1: Alignment Regimes

Noisiness in $\mathcal{D}$ is partly responsible for alignment resistance Finally, we investigate which features predict alignment resistance by running the following logistic regression:

$$\log\left(\frac{p_\delta}{1-p_\delta}\right) = \alpha_0 + \sum_{\tau\in\mathcal{T}} \rho^c(\tau)t_i^c(\tau) + \rho^r(\tau)t_i^r(\tau) + \varepsilon_i, \tag{3}$$

where $p_\delta = \Pr[\delta_i = 1]$ represents the probability of alignment, and $\rho^*(\tau)$ are the regression coefficients. All else being equal, eloquent entries increase the odds of misalignment by $\exp(\rho^c(\text{eloquence}))$.

In Figure 3 (right), we observe that chosen entries exhibiting positive features (e.g., positivity, eloquence, harmlessness, helpfulness) and rejected entries exhibiting negative features (e.g., sexually explicit content, breaking privacy) reduce the likelihood of misalignment. Conversely, chosen entries exhibiting negative features and rejected entries exhibiting positive features increase misalignment. These estimates are consistent with the observations in Figure 3 (center). Recall from Figure 2 that most rewards are in the third quadrant (around $(-1,-1)$) and most reward shifts are small. In such cases, a positive θ_i is more likely to convert a misaligned reward vector pre-$\mathcal{D}$ to an aligned reward vector post-$\mathcal{D}$ and, conversely, a negative θ_i is more likely to convert an aligned reward vector to a misaligned reward vector. For most features, this association holds: for instance, harmlessness in rejected entries is associated with a negative θ_i in Figure 3 (center) and with increased misalignment in Figure 3 (right). Similar patterns are observed for refusal, sexually explicit content, breaking privacy, and sentiment.[11]

[11] Interestingly, the relationship between reward shifts and misalignment is sometimes reversed. For example, a helpful rejected entry leads to both a positive reward shift and increased misalignment (compared to a non-helpful one). Similarly, eloquence in chosen entries leads to a negative reward shift and reduced misalignment. A similar pattern is observed for manipulation in chosen entries, though the reward shifts are not statistically significantly positive in that case. These observations suggest that some relevant reward vectors may be closer to the $(1,1)$ point in the first quadrant and may become misaligned through positive reward shifts.

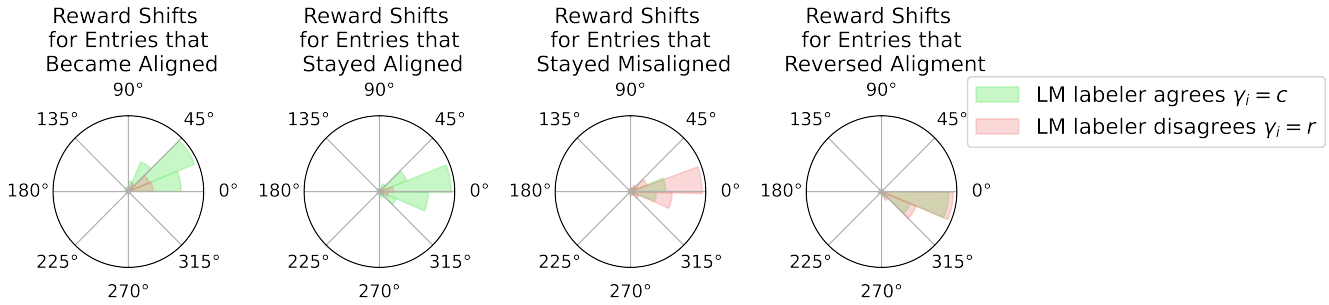

Figure 4: Reward shifts broken down by LM-labeler preference (green for $\gamma_i = c$ and pink for $\gamma_i = r$). Each column corresponds to a different alignment regime, from left to right: pairs that became aligned ($(1 - \underline{\delta_i})\delta_i = 1$), that remained aligned ($\underline{\delta_i}\delta_i = 1$), that resisted alignment ($(1 - \underline{\delta_i})(1 - \delta_i) = 1$), and reversed alignment ($\underline{\delta_i}(1 - \delta_i) = 1$).

These findings suggest that the RM predominantly learns desirable features, with misalignment partly arising when rejected entries are too "good" (e.g., too eloquent or harmless) or chosen entries are too "bad" (e.g., sexually explicit or manipulative). Additionally, misalignment can occur when chosen and rejected entries are too similar overall, indicating that the lack of a strong distinction between these entries contributes to misalignment. This finding could indicate either that the human comparisons over these entries are likely to be noisy or the RM is not sufficiently accurate to distinguish between these types of entries. However, this analysis does not address cases where spoiler features conflict with target features and mislead the RM, a topic we explore in the next section on alignment robustness.

2.3 How do mild perturbations in entries' features change the RM's alignment?

This section examines the robustness of feature imprinting in the post-$\mathcal{D}$ RM $\mathcal{R}$ through mild perturbations.

Robustness Scores We employ an LM-rewriter to modify a subset of the paired entries of the alignment dataset, adjusting the stylistic tone while preserving the original meaning. We control for changes in semantic meaning using cosine similarity between vectors generated by a text similarity model between the original and rewritten entries. We denote any rewritten entity (e.g., t, $\mathcal{D}$, δ) with a hat symbol (e.g., $\widehat{t}$). The robustness score is computed as the coefficient of a logistic regression that measures the impact of label flipping on misalignment incidence. The indicator variable $\delta_i(1-\widehat{\delta_i})$ equals 1 when the RM was aligned with human preferences before rewriting and not after. We estimate the robustness scores π^* as follows:

$$\log\left(\frac{\widehat{p_\delta}}{1-\widehat{p_\delta}}\right) = \alpha_0 + \sum_{\tau\in\mathcal{T}} \pi^*(\tau)\left(t_i^*(\tau) - \widehat{t_i^*(\tau)}\right) + \varepsilon_i. \tag{4}$$

where $\widehat{p_\delta} = \Pr[\delta_i(1 - \widehat{\delta_i}) = 1]$ represents the probability of misalignment after rewriting, and $t_i^*(\tau) - \widehat{t_i^*(\tau)}$ is a categorical variable that can take values in $-1, 0, 1$. We set 0 (the absence of label flip) as the baseline, resulting in two coefficients $\pi^*(\tau)$, denoted $\pi_+^*(\tau)$ and $\pi_-^*(\tau)$. For example, $\pi_-^c(\tau) > 0$ indicates that a chosen entry becoming more eloquent increases the likelihood of misalignment. Specifically, $\pi_-^c(\text{eloquent})$ is interpreted as follows: pairs where the chosen entry becomes more eloquent after rewriting have $\exp(\pi_-^c(\text{eloquent}))$ times higher odds of misalignment compared to pairs without such flips. Similarly, pairs where the rejected entry becomes less eloquent after rewriting lead to $\exp(\pi_+^r(\text{eloquent}))$ times higher odds of misalignment than pairs without such flips. Thus, $\pi_*^*(\tau)$ measures the extent to which alignment is robust to rewriting, isolating the effects of each feature and each event type.

Rewriting caused more misalignment due to shifts in texts' positivity We perform surface-level rewriting of a random 1% subset of $\mathcal{D}$ with Mistral 7B v0.1 Instruct[12]. The rewritten dataset was then featurized, focusing on the following features: helpfulness, harmlessness, coherence, eloquence, and sentiment. Our analysis concentrated on entries where the helpfulness and harmlessness labels remained unchanged after rewriting, filtering out potential sensitivity of the LM-labeler to the rewriting process.[13]

The alignment score on rewritten entries is $\widehat{a_+} = 0.71$, indicating a $3-$point drop in alignment due to rewriting. An analysis of the results of Eq. (4) displayed in Figure 5, reveals that only the robustness scores $\pi_+^c(\text{sentiment})$ and $\pi_-^r(\text{sentiment})$ are statistically significant. All else being equal, when a chosen entry becomes less positive after rewriting, the odds of misalignment are multiplied by $\exp(\pi_+^c(\text{sentiment})) = \exp(0.12) = 1.13$ compared to cases without rewriting-induced label flips. Similarly, when a chosen entry becomes more positive after rewriting, the odds of misalignment are multiplied by $\exp(\pi_-^r(\text{sentiment})) = 1.12$ compared to entries without rewriting-induced label flips.

[12]Rewriting was performed with the prompt in Appendix C.5 using an FP16 version of Mistral 7B ran at temperature 0.1 via Ollama. Output format was controlled using Ollama's JSON mode. We use BGE-m3 (Multi-Granularity 2024), a general-purpose text-similarity model, to measure cosine similarity; see Appendix D.7.

[13]See Appendix D.6 for a distribution of the feature flips.

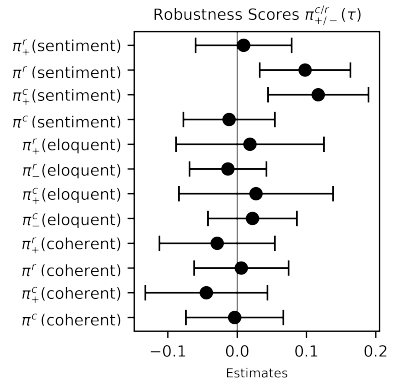

Figure 5: Robustness scores $\pi^{c/r}_{+/-}(\tau)$ across entry types (c or r), contrasts ($+$ or $-$) and features τ.

3 Discussion

We (a) evaluate how well RMs learn desired behaviors like harmlessness and spoiler features, (b) identify reasons for persistent alignment resistance after training, and (c) assess the impact of minor dataset perturbations on feature imprint stability. Testing our method on the Anthropic/hh-rlhf preference dataset and OpenAssistant RMs that while alignment improves rewards for desirable traits and penalties for harmful content, significant misalignment with human preferences persists. Alignment resistance may stem from several sources: (i) concept confusion within $\mathcal{D}$, (ii) inconsistencies between $\mathcal{D}$ and the RM's training datasets, and (iii) discrepancies between the RM and its base training data.

Notably, 73% of $\mathcal{D}'$s entries have $\gamma_i = i$, suggesting many entries in a pair are difficult to differentiate by the LM-labeler. Appendix D.1 further shows that the rewards assigned to each of the paired entries are remarkably similar (see Figure 8's bright diagonal) and manual assessments confirm entries are often indistinguishable. Also, Section 2.2 indicates that "good" rejected entries and "bad" chosen entries contribute to misalignment, suggesting the RM may correctly reward desirable features present in rejected entries (and vice versa). This could increase the misalignment incidence, as small perturbations in a reward vector close to the diagonal can tip it from aligned to misaligned. These findings support hypothesis (i) on concept confusion within $\mathcal{D}$ as a significant contributor to fine-tuning failures.

The lack of robustness to certain spoiler features also indicates that the RM may sometimes reward the wrong features, supporting hypothesis (iii) on concept confusion between the RM and its base training data. Regarding hypothesis (ii), an RM, as a pre-trained language model, begins with an initial semantic representation based on its pre-training data, which is reshaped during retraining. We posit that the LM-labeler's agreement with the RM on alignment resistance suggests a shared latent representation of these features. This observation may indicate a relationship between the compositions of the pre-training and fine-tuning data. However, without access to the pre-training data, we cannot test this hypothesis directly.

Limitations Our methodology depends on the taxonomy labels used to evaluate alignment. Robustness checks in Appendix D.5 indicate that some labels may be unstable when assessed by different LM-labelers. Although we believe these labels are at least as reliable as human labels (Gilardi, Alizadeh, and Kubli 2023), the issue of label quality is not unique to our study and requires ongoing scrutiny to avoid circularity when using LMs to assess LM alignment.

Additionally, our approach does not systematically identify and define different "spoiler" features. While some features may be universally applicable across various pipelines, specific contexts might necessitate the development of more tailored frameworks to accurately detect and address potential confounding factors in RM behaviors. Future work should focus on identifying and managing these features to enhance the efficacy of alignment pipelines.

Systematic error analyses are also needed to explore how various elements of the alignment pipeline interact. This work examines the interconnections between an alignment dataset and a series of RMs as a first step in this direction. High-quality taxonomy labels could accompany the entries of the alignment dataset alongside human or synthetic preferences. These labels would help ensure that spoiler features are balanced across value targets and that human preferences are internally consistent. They would also provide a priori and testable objectives for feature imprint, enabling rigorous measurement and mitigation of the impact of spoiler features through additional training.

Future work The pre-$\mathcal{D}$ RM was trained on a corpus of three semantic datasets (web-gpt, summarize from feedback, and synthetic-instruct-gptj-pairwise) designed to train RMs on semantic tasks. Resistance to alignment on these tasks is also observed and can be studied using our proposed method (resistance incidences of 49% and 66% are observed with web-gpt and summarize from feedback, respectively).[14]

Next, the importance of having a high-quality alignment pipeline becomes paramount as powerful base models are open-sourced. To the best of our knowledge, the combination of the Anthropic/hh-rlhf alignment dataset and the OpenAssistant RMs are among the most popular alignment tools on Hugging Face and they were crucial for improving our understanding of alignment dynamics in this work. We hope that such efforts will support the development of even better open-source alignment pipelines, and we would be excited about new research that releases and scrutinizes both datasets and openly shared RMs.

In conclusion, we posit that alignment datasets and RMs are crucial for providing granular interpretations of value alignment. We developed a methodology to test the performance of RMs relative to their training alignment dataset and value objectives. We hope the paper raises awareness of these issues and introduces a first generation of evaluation metrics.

[14]See the numbers reported by OpenAssistant on the reward-model-deberta-v3-large-v2 page. The small discrepancy between our computation and theirs appears to be due to OpenAssistant's tokenization procedure to save compute space.

Acknowledgments

We thank Benjamin Steinberg, Jack Cushman, Jenn Louie, Jonathan Zittrain, Miles Brundage, Naomi Bashkansky, Neil Shah, Tom Zick and Zhi Rui Tam for their useful advice. This work was conducted through OpenAI's Researcher Access Program (https://openai.com/form/researcher-access-program/) to allow the permissive queries required for the study and supported through that program's API credits.

References

Ahmadian, A.; Cremer, C.; Gallé, M.; Fadaee, M.; Kreutzer, J.; Üstün, A.; and Hooker, S. 2024a. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *arXiv preprint arXiv:2402.14740*.

Ahmadian, A.; Ermis, B.; Goldfarb-Tarrant, S.; Kreutzer, J.; Fadaee, M.; Hooker, S.; et al. 2024b. The Multilingual Alignment Prism: Aligning Global and Local Preferences to Reduce Harm. *arXiv preprint arXiv:2406.18682*.

Alex Havrilla. 2023. synthetic-instruct-gptj-pairwise (Revision cc92d8d).

Askell, A.; Bai, Y.; Chen, A.; Drain, D.; Ganguli, D.; Henighan, T.; Jones, A.; Joseph, N.; Mann, B.; DasSarma, N.; et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Bansal, H.; Dang, J.; and Grover, A. 2023. Peering through preferences: Unraveling feedback acquisition for aligning large language models. *arXiv preprint arXiv:2308.15812*.

Cahyawijaya, S.; Chen, D.; Bang, Y.; Khalatbari, L.; Wilie, B.; Ji, Z.; Ishii, E.; and Fung, P. 2024. High-Dimension Human Value Representation in Large Language Models. *arXiv preprint arXiv:2404.07900*.

Casper, S.; Davies, X.; Shi, C.; Gilbert, T. K.; Scheurer, J.; Rando, J.; Freedman, R.; Korbak, T.; Lindner, D.; Freire, P.; et al. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Conitzer, V.; Freedman, R.; Heitzig, J.; Holliday, W. H.; Jacobs, B. M.; Lambert, N.; Mossé, M.; Pacuit, E.; Russell, S.; Schoelkopf, H.; et al. 2024. Social choice for AI alignment: Dealing with diverse human feedback. *arXiv preprint arXiv:2404.10271*.

Dubois, Y.; Li, C. X.; Taori, R.; Zhang, T.; Gulrajani, I.; Ba, J.; Guestrin, C.; Liang, P. S.; and Hashimoto, T. B. 2024. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36.

Findeis, A.; Kaufmann, T.; Hüllermeier, E.; Albanie, S.; and Mullins, R. 2024. Inverse Constitutional AI: Compressing Preferences into Principles. *arXiv preprint arXiv:2406.06560*.

Ganguli, D.; Lovitt, L.; Kernion, J.; Askell, A.; Bai, Y.; Kadavath, S.; Mann, B.; Perez, E.; Schiefer, N.; Ndousse, K.; et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.

Gao, L.; Schulman, J.; and Hilton, J. 2023. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, 10835–10866. PMLR.

Ge, L.; Halpern, D.; Micha, E.; Procaccia, A. D.; Shapira, I.; Vorobeychik, Y.; and Wu, J. 2024. Axioms for AI Alignment from Human Feedback. *arXiv preprint arXiv:2405.14758*.

Gilardi, F.; Alizadeh, M.; and Kubli, M. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30): e2305016120.

Gilbert, T. K.; Lambert, N.; Dean, S.; Zick, T.; Snoswell, A.; and Mehta, S. 2023. Reward reports for reinforcement learning. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 84–130.

Glaese, A.; McAleese, N.; Trebacz, M.; Aslanides, J.; Firoiu, V.; Ewalds, T.; Rauh, M.; Weidinger, L.; Chadwick, M.; Thacker, P.; et al. 2022. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*.

He, P.; Gao, J.; and Chen, W. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Hosking, T.; Blunsom, P.; and Bartolo, M. 2023. Human feedback is not gold standard. *arXiv preprint arXiv:2309.16349*.

Kirk, H. R.; Whitefield, A.; Röttger, P.; Bean, A.; Margatina, K.; Ciro, J.; Mosquera, R.; Bartolo, M.; Williams, A.; He, H.; et al. 2024. The PRISM Alignment Project: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models. *arXiv preprint arXiv:2404.16019*.

Lambert, N.; and Calandra, R. 2023. The alignment ceiling: Objective mismatch in reinforcement learning from human feedback. *arXiv preprint arXiv:2311.00168*.

Lambert, N.; Krendl Gilbert, T.; and Zick, T. 2023. The history and risks of reinforcement learning and human feedback. *arXiv e-prints*, arXiv–2310.

Lambert, N.; Pyatkin, V.; Morrison, J.; Miranda, L.; Lin, B. Y.; Chandu, K.; Dziri, N.; Kumar, S.; Zick, T.; Choi, Y.; et al. 2024. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*.

Li, X.; Zhang, T.; Dubois, Y.; Taori, R.; Gulrajani, I.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. Alpacaeval: An automatic evaluator of instruction-following models.

Liang, P.; Bommasani, R.; Lee, T.; Tsipras, D.; Soylu, D.; Yasunaga, M.; Zhang, Y.; Narayanan, D.; Wu, Y.; Kumar, A.; et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Multi-Granularity, M.-L. M.-F. 2024. M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. *OpenReview*.

Nakano, R.; Hilton, J.; Balaji, S.; Wu, J.; Ouyang, L.; Kim, C.; Hesse, C.; Jain, S.; Kosaraju, V.; Saunders, W.; et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Niu, Z.; Ren, H.; Gao, X.; Hua, G.; and Jin, R. 2024. Jailbreaking attack against multimodal large language model. *arXiv preprint arXiv:2402.02309*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; and Christiano, P. F. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33: 3008–3021.

Swayamdipta, S.; Schwartz, R.; Lourie, N.; Wang, Y.; Hajishirzi, H.; Smith, N. A.; and Choi, Y. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. *arXiv preprint arXiv:2009.10795*.

Wang, B.; Zheng, R.; Chen, L.; Liu, Y.; Dou, S.; Huang, C.; Shen, W.; Jin, S.; Zhou, E.; Shi, C.; et al. 2024. Secrets of rlhf in large language models part ii: Reward modeling. *arXiv preprint arXiv:2401.06080*.

Wirth, C.; Akrour, R.; Neumann, G.; and Fürnkranz, J. 2017. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 18(136): 1–46.

Wu, M.; and Aji, A. F. 2023. Style over substance: Evaluation biases for large language models. *arXiv preprint arXiv:2307.03025*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Zheng, R.; Dou, S.; Gao, S.; Hua, Y.; Shen, W.; Wang, B.; Liu, Y.; Jin, S.; Liu, Q.; Zhou, Y.; et al. 2023. Secrets of rlhf in large language models part i: Ppo. *arXiv preprint arXiv:2307.04964*.

Ziegler, D. M.; Stiennon, N.; Wu, J.; Brown, T. B.; Radford, A.; Amodei, D.; Christiano, P.; and Irving, G. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.