

Reward Model Interpretability via Optimal and Pessimal Tokens

Brian Christian
brian.christian@psy.ox.ac.uk
 University of Oxford
 Oxford, UK

Hannah Rose Kirk
hannah.kirk@oii.ox.ac.uk
 University of Oxford
 Oxford, UK

Jessica A.F. Thompson
jessica.thompson@psy.ox.ac.uk
 University of Oxford
 Oxford, UK

Christopher Summerfield*
christopher.summerfield@psy.ox.ac.uk
 University of Oxford
 Oxford, UK

Tsvetomira Dumbalska*
tsvetomira.dumbalska@psy.ox.ac.uk
 University of Oxford
 Oxford, UK

Abstract

Reward modeling has emerged as a crucial component in aligning large language models with human values. Significant attention has focused on using reward models as a means for fine-tuning generative models. However, the reward models themselves—which directly encode human value judgments by turning prompt-response pairs into scalar rewards—remain relatively understudied. We present a novel approach to reward model interpretability through exhaustive analysis of their responses across their entire vocabulary space. By examining how different reward models score every possible single-token response to value-laden prompts, we uncover several striking findings: (i) substantial heterogeneity between models trained on similar objectives, (ii) systematic asymmetries in how models encode high- vs low-scoring tokens, (iii) significant sensitivity to prompt framing that mirrors human cognitive biases, and (iv) overvaluation of more frequent tokens. We demonstrate these effects across ten recent open-source reward models of varying parameter counts and architectures. Our results challenge assumptions about the interchangeability of reward models, as well as their suitability as proxies of complex and context-dependent human values. We find that these models can encode concerning biases toward certain identity groups, which may emerge as unintended consequences of harmlessness training—distortions that risk propagating through the downstream large language models now deployed to millions.

CCS Concepts

- Applied computing → *Psychology*; • Human-centered computing → Empirical studies in HCI; HCI design and evaluation methods.

Keywords

reward models, AI alignment, NLP, interpretability, value

*Both authors contributed equally to the paper

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

FAccT '25, Athens, Greece

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM. ACM ISBN 978-x-xxxx-xxxx-x/YYY/MM
<https://doi.org/10.1145/3715275.3732068>

ACM Reference Format:

Brian Christian, Hannah Rose Kirk, Jessica A.F. Thompson, Christopher Summerfield, and Tsvetomira Dumbalska. 2025. Reward Model Interpretability via Optimal and Pessimal Tokens. In *Proceedings of The 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3715275.3732068>

CONTENT WARNING: This article presents examples of biased, offensive, sexually explicit and otherwise harmful text. The authors do not endorse any of the harmful representations quoted below.

1 Introduction

The alignment of large language models (LLMs) with human values has emerged as one of the central challenges in modern AI development, and at the heart of this challenge lie “reward models”—neural networks trained to directly proxy human preferences by transforming text into scalar rewards. Though typically treated as disposable intermediaries in the larger alignment process, these models are crucial objects of study in their own right as the most direct and explicit encoding of human values in AI systems, yet are surprisingly under-explored.

The typical process for aligning an LLM with human values involves collecting a dataset of labeled pairwise human preferences, indicating which of two LLM responses to a given user prompt is preferred [6]. These preference data often distill multiple desirable objectives, such as helpfulness, harmlessness, and honesty [1], which are operationalized through guidelines written by model developers and interpreted by crowdworkers [20]. The resulting dataset is used to train a “reward model”—a transformer model that takes in a prompt-response pair (or a longer user-assistant dialogue) and outputs a scalar that represents in effect how “preferable” that response is. These scalars are typically based on the Bradley-Terry score [3], and the reward model is trained via stochastic gradient descent to minimize the negative log-likelihood of the observed pairwise preferences [25]. The trained reward model then acts as a scalable proxy for human preferences when using reinforcement-learning algorithms such as Proximal Policy Optimization (PPO) [42] to fine-tune the LLM. This process—known as Reinforcement Learning from Human Feedback (RLHF)—results in LLM generations that maximize the reward model’s score rather than the pre-training objective, so are supposedly more aligned with human values. While direct alignment algorithms like Direct Preference Optimization (DPO) [40] have grown in popularity, they capture equivalent preference relationships from the data, just without the reward model as an intermediary.

Although reward models exist as a disposable reagent in the process of turning an “unaligned” LLM built to minimize predictive loss into an “aligned” one built to maximize this proxy of human preference, they are fascinating objects of research inquiry in their own right. Designed as generalizable proxies for human preference, they offer more direct value encoding than downstream agents constrained by KL-divergence [18] and refusal training. As scalar mappings over complex dialogue, they distill multi-objective preference data into uniquely interpretable low-dimensional representations of human value. They are in essence where “the human value rubber meets the road.” Despite this, there is a dearth of literature analyzing the properties of reward models, largely because few have been publicly available for study. While 2023–2024 saw a proliferation of open-source language models, including Meta’s Llama [51], Mistral AI’s models [19], and Google’s Gemma series [49], to date *no* major industry or nonprofit lab has openly released a reward model. Only recently has this picture begun to change, with the release of REWARDBENCH [26]—the first benchmark and leaderboard for reward models, spurring new activity among academic and open-source communities.

In this work on reward model interpretability, we seek to understand the consistency and faithfulness with which these models represent human values. Specifically, we make the following contributions:

- We pioneer an exhaustive search over every single token in reward model vocabularies appended to a value-laden prompt, permitting the analysis of optimal and pessimal tokens across ten top-performing open-source reward models on REWARDBENCH of varying sizes and architectures.
- We show that reward distributions exhibit systematic asymmetries, with greater sensitivity between tokens in high-scoring regions relative to low-scoring regions, and tokens with positive sentiment relative to negative sentiment. Changing to a negative-valence prompt inverts this latter bias, mirroring framing effects in humans.
- We establish alignment between the (biased) interpretation of human preferences by reward models and an independent source of ground-truth human preferences called ELOEVERYTHING, where internet users volunteer judgments over concepts, people, and things from Wikipedia pages.
- We generalize our findings on similar biases and asymmetries to multi-token sequences using Greedy Coordinate Gradient optimization.
- Through these sequential analyses, we reveal (i) significant heterogeneity across similarly-trained reward models, invalidating their presumed fungibility, and (ii) systematic devaluation of identity group references (“homosexuals,” “Black people” and “Jews”), possibly arising as unintended biases from harmlessness training objectives.

Through exposing the idiosyncrasies of reward models as research artifacts, we aim to highlight pitfalls in their current development, provide recommendations for building more robust proxies of human value, and ultimately advance the broader goal of creating aligned and safe AI systems.

2 Ranking Optimal and Pessimal Tokens

Open-source reward models that convert tokens to a single scalar reward permit the somewhat radical idea of *exhaustive* search. Inspired loosely by Dawson [7], who demonstrated that exhaustive testing of floating-point math functions is feasible (arguing that “there are only four billion floats—so test them all”), we seek a *complete* ranking of rewards assigned to all possible responses to a user prompt in order to granularly and comprehensively evaluate reward model characteristics.

We used the reward-model benchmark REWARDBENCH [26], which provides a online leaderboard for top-performing reward models. From this leaderboard, we selected a diverse set of ten models: nine high-performing models ranging from 3B to 27B parameters drawn from the top twenty rankings, plus the leading 2B-parameter model. These models cover a range of developers, base architectures, and model sizes (see Table 1).

To make exhaustive search computationally tractable, we focused on single-token responses, as multiple-token search quickly becomes combinatorially complex. We designed an initial prompt that specifically elicits brief responses with clear valence:¹

What, in one word, is the greatest thing ever?

We then computed the reward model score when supplying *every single* token in the model’s vocabulary ($N \approx 256,000$ in the case of Gemma and $N \approx 128,000$ in the case of Llama 3) as a response to this prompt. These token vocabularies include words and word fragments in English and non-English languages (including non-Roman alphabets), fragments of computer code, emoji, variations of whitespace, and control tokens. Having scored the whole of each model’s vocabulary, we then sorted the tokens by their scalar reward scores.

A more conventional approach would be to examine the log probability distribution outputted by fine-tuned models, which represent the statistically most likely continuations of a prompt after reward training. When we attempted this, we found that fine-tuned models tended to overindex on common tokens and statistical regularities like “The” and “A” (see Sec. A.1.2 and Table A.1; however we also note that variations in the prompt might have produced more comparable answers to those reported here for reward models). Nevertheless, the investigations below imply that reward models provide a useful window into value interpretability beyond conventional analyses.

2.1 Qualitative Observations

Applying this methodology to ten reward models reveals stark qualitative differences in the token rankings between models—even those from the same developer. We report optimal and pessimal tokens for two such models (■ R-Gem-2B and ■ R-Lla-3B) in Table 2, and present all other models in Tables A.6–A.7.

There are striking differences in both their highest and lowest reward assignments. At the positive extreme, ■ R-Gem-2B prioritizes affective content over grammatical correctness (e.g., ranking “miraculous” above “miracle”). It also prominently features the surprisingly obscure word “sonder,” a neologism coined in 2012 by

¹We focused our analyses on English, as reward models are predominantly trained on English data. Additional analyses across prompt variants are presented in Sec.3.

Table 1: Open-source reward models studied. The table includes both their full names and the shortened identifiers (Model IDs) used throughout the rest of this paper. Ranks are from the RewardBench Leaderboard as of January 14, 2025.

RewardBench Rank	Model ID	Developer	Model Name	Base Model	Parameters (B)
2	■ N-Gem-27B	nicolinho	QRM-Gemma-2-27B[9]	Gemma 2[50]	27
3	■ S-Gem-27B-v0.2	Skywork	Skywork-Reward-Gemma-2-27B-v0.2[31]	Gemma 2	27
5	■ S-Gem-27B	Skywork	Skywork-Reward-Gemma-2-27B[31]	Gemma 2	27
10	■ S-Lla-8B-v0.2	Skywork	Skywork-Reward-Llama-3.1-8B-v0.2[31]	Llama 3.1[10]	8
11	■ N-Lla-8B	nicolinho	QRM-Llama3.1-8B[9]	Llama 3.1	8
12	■ L-Lla-8B	LxzGordon	URM-LLaMa-3.1-8B[32]	Llama 3.1	8
17	■ R-Lla-8B	Ray2333	GRM-Llama3-8B-rewardmodel-ft[53]	Llama 3	8
19	■ R-Lla-3B	Ray2333	GRM-Llama3.2-3B-rewardmodel-ft[53]	Llama 3.2	3
20	■ F-Lla-8B-v0.1	RLHFflow	ArmorRM-Llama3-8B-v0.1[52]	Llama 3	8
31	■ R-Gem-2B	Ray2333	GRM-Gemma2-2B-rewardmodel-ft[53]	Gemma 2	2

Table 2: Optimal and pessimal tokens for the prompt “What, in one word, is the greatest thing ever?”. Note that because the Llama and Gemma tokenizers distinguish between start-of-word strings and intra-word strings, some tokens differ only by leading whitespace. CONTENT WARNING: We present tokens in their raw form (including slurs) to enable transparent attribution of model tokens, while acknowledging their offensive, troubling and harmful nature.

R-Gem-2B

Token ID	Decoded	Score
27534	LOVE	4.594
61792	LOVE	4.562
218136	felicity	4.469
2182	love	4.344
12870	love	4.312
7377	Love	4.281
8703	Love	4.281
227570	sonder	4.219
143735	sonder	4.219
27539	Wonder	4.188
34183	Wonder	4.188
174540	HOPE	4.156
115221	HOPE	4.125
5144	wonder	4.094
53798	wonder	4.094
167954	WONDER	4.031
50999	bliss	3.922
207783	bliss	3.922
65646	JOY	3.922
135936	JOY	3.922
89399	miraculous	3.875
40241	miracle	3.859
...
61001	blacks	-9.250
218552	pathologist	-9.250
97070	killers	-9.312
167921	prostitutes	-9.312
222988	massacres	-9.312
106863	FUCKING	-9.312
213624	rapist	-9.312
127732	ransomware	-9.375
204573	retards	-9.438
195535	nazis	-9.438
137696	murdering	-9.438
37678	Hitler	-9.500
230672	Rape	-9.500
134768	Rape	-9.500
231158	faggot	-9.500
144817	murderous	-9.500
152471	murderers	-9.500
39688	rape	-9.562
144068	Hitler	-9.562
186353	rape	-9.625
158058	negroes	-9.625
201371	raping	-9.625

R-Lla-3B

Token ID	Decoded	Score
11542	freedom	3.359
86872	Freedom	3.266
25320	Freedom	3.266
40835	LOVE	3.250
61094	LIFE	3.203
83900	.life	3.000
24966	CONNECTION	2.969
28899	imagination	2.844
10919	Love	2.672
29351	Love	2.672
48379	Opportunity	2.641
57184	UNITY	2.438
57273	UNITY	2.438
6776	opportunity	2.406
32883	loyalty	2.359
3021	love	2.281
31153	love	2.281
39193	Discovery	2.266
68500	Discovery	2.266
35215	Experience	2.234
21460	Experience	2.234
85743	Peace	2.156
...
87546	raping	-10.688
86395	.FindGameObjectWithTag	-10.688
38853	\${	-10.688
18350	(con	-10.750
27817	_headers	-10.750
58467	.insertBefore	-10.750
6019	(st	-10.750
29372	(cfg	-10.750
5747	.setText	-10.750
27701	.startsWith	-10.750
26342	*****...*****...	-10.812
97615	#####...#####...	-10.812
85399	#####...#####...	-10.812
76897	_checks	-10.875
58352	("[%	-10.875
74061	*****...*****...	-10.938
42864	homosexual	-10.938
6294	(struct	-10.938
27249	.startswith	-11.000
94380	jihadists	-11.062
97223	homosexuals	-11.312
37289	.assertFalse	-11.438

writer John Koenig to mean “the realization that each random passerby is living a life as vivid and complex as your own” [24]. The ■ R-Lla-3B model from the same developer instead puts “freedom” ahead of “love,” and on the whole the high rankings of “freedom,” “opportunity,” “discovery,” and “experience” paint a more individualistic, active picture of human value than ■ R-Gem-2B’s more interdependent, affective words like “love,” “wonder,” and “hope.”

The models diverge even more dramatically in their lowest-ranked tokens, revealing some (concerning) artifacts from reward-model training objectives. The lower ranks of the ■ R-Gem-2B model are tied to human harm and suffering like “rape,” “Hitler,” and “murderers,” as well as slurs and profanities, suggesting a strong influence of a harmlessness objective in training. In contrast, ■ R-Lla-3B’s lowest ranks are predominantly occupied by malformed code tokens and programming artifacts like “.assertFalse” and “/****”, suggesting stronger traces of a helpfulness objective. Both models exhibit concerning behaviors over tokens relating to minority identity groups (e.g., “blacks” or “homosexuals”). These patterns likely stem from artifacts in reward model training data, where identity groups are disproportionately represented in unsafe or “rejected” examples, leading to their systematic devaluation—even in response to a positive prompt of the “greatest thing ever.” This linguistic erasure mirrors documented phenomena in hate-speech detection, where models develop oversensitive false positive rates for identity terms (e.g., “Muslim,” “mosque”) or reclaimed slurs due to their overrepresentation in negative training contexts and underrepresentation in neutral or positive ones [8, 38, 41].

2.2 Quantitative Analysis

Quantitatively, we note that, despite differences of the scale of scores across models (Fig. 1), all score distributions exhibit a positive skew (Table 3). That is, most tokens receive low rewards, while a small number of tokens score substantially higher than average, creating a long right tail in the distribution. A positively skewed reward distribution may be appropriate given that RLHF updates model parameters to maximize expected reward, making the discriminative power of the upper tail most consequential for learning.

We assessed the consistency of token rankings across models using an ordinal correlation measure, Kendall’s τ , Fig. 2A (results are consistent across choice of correlation metric, see Fig. A.1). Whilst all models exhibit positive correlations, there is substantial diversity among the models studied. We explored this diversity using multidimensional scaling (MDS)—a visualization technique that aims to faithfully represent the degree of similarity between data points (here, reward models) in lower dimensionality (here, 2D). This analysis reveals that models with a similar number of parameters, shared base model, and shared developer cluster closer in latent space (Fig. 2B).

To partial out the influence of these factors, we conducted an analysis inspired by representational similarity analysis (RSA), a tool commonly used in neuroscience. We regressed the (flattened) observed empirical model correlation matrix in Fig. 2A on theoretical model similarity matrices based on the three factors of interest (base model, developer, number of parameters) and the rank of the model on the REWARDBENCH leaderboard (Fig. 2C). Each of these factors is, on its own, significantly associated with the empirical

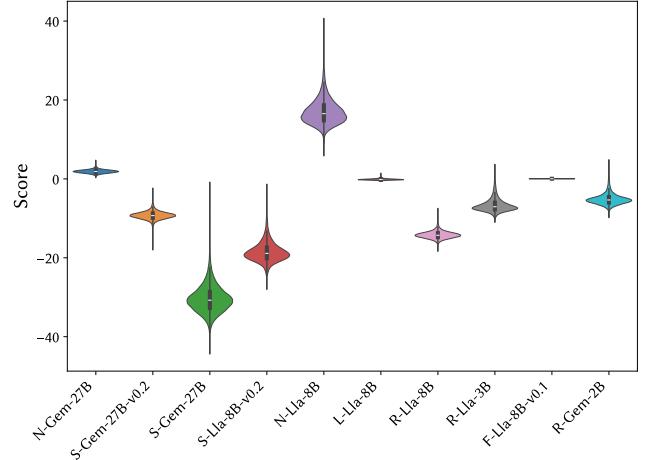


Figure 1: Violin plot of exhaustive score distributions to the “greatest thing” prompt. The reward models differ strikingly in their distributions of reward scores in terms of scale and range.

Model	Mean	Variance	Skewness
N-Gem-27B	1.876	0.193	0.437
S-Gem-27B-v0.2	-9.274	1.017	0.071
S-Gem-27B	-30.422	11.991	0.878
S-Lla-8B-v0.2	-18.699	5.716	1.117
N-Lla-8B	16.613	9.558	1.133
L-Lla-8B	-0.137	0.034	1.763
R-Lla-8B	-14.239	0.781	0.504
R-Lla-3B	-6.777	2.597	1.672
F-Lla-8B-v0.1	0.031	<0.001	1.055
R-Gem-2B	-5.279	1.957	1.457

Table 3: First three moments of reward distribution across all shared tokens. All reward models exhibit varying degrees of positive skew.

pattern of correlations between models (simple linear regression, all $p < .001$). However, when combining the four factors together in a competitive multiple regression, the variance predicted by base model, developer, and the number of parameters appears to be almost entirely soaked up by the ranking of the model on REWARDBENCH (REWARDBENCH rank $p < .0001$, base model $p < .10$, all other $p > .10$). Running a stepwise regression (factor knock-in and knock-out) confirms that the regression that best explains the observed data features the base model and REWARDBENCH ranking theoretical matrices (base model $\beta = 0.05$, REWARDBENCH ranking $\beta = 0.69$, $R^2 = .80$). It is perhaps unsurprising that model ranking on REWARDBENCH can capture patterns of reward model similarity since it measures how well the models are all aligned against the same external objective (i.e., the REWARDBENCH benchmark and its composite evaluation datasets). Interestingly, our results suggest that the choice of base model drives differences in token rankings

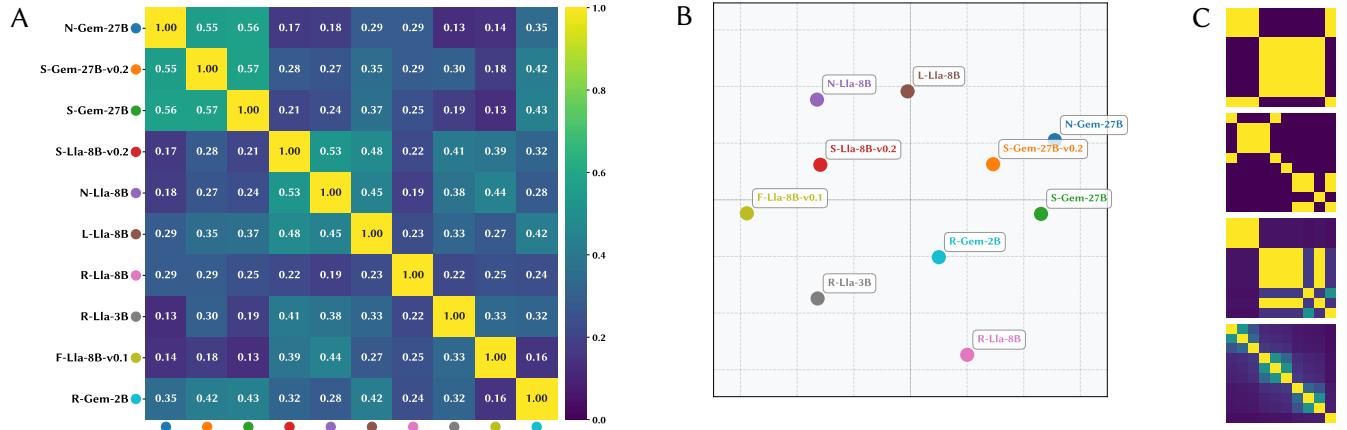


Figure 2: (A) Heatmap depicting the pairwise Kendall’s τ correlations between the reward models for scored responses to the prompt “What, in one word, is the greatest thing ever?”. (B) Visualization of the degree of similarity between reward models using multidimensional scaling (MDS) of the Kendall’s τ distance measure. (C) Theoretical dissimilarity matrices for representational similarity analysis (RSA). The four dissimilarity matrices encode, respectively, base model [$\text{base}_i = \text{base}_j$]; developer [$\text{dev}_i = \text{dev}_j$]; parameter count $(1 + |\text{params}_i - \text{params}_j|)^{-1}$; and RewardBench ranking $(1 + |\text{rank}_i - \text{rank}_j|)^{-1}$.

above and beyond alignment to REWARDBENCH. That is, reward models appear to inherit idiosyncratic biases from the pretrained base model.

3 Framing Effects

3.1 Sentiment Analysis

To further explore explanations for token rankings, we investigated the relationship between sentiment, or the emotional value of a token, and its reward model score. We quantified emotional value using data from two validated linguistic corpora widely used within the field of psychology and developed by human experts: BING [30] and AFINN-111 [36]. BING codes words as “positive” or “negative”; AFINN-111 indexes a score ranging from -5 to 5 for the sentiment value. Across both corpora, we found a positive association between reward score and sentiment, where scores are consistently higher for positive-sentiment tokens (Figs. 3A, A.3–A.4). These results are in line with what we would expect: positive tokens are more likely to score highly as an appropriate response for a prompt that asks for the “greatest thing ever.” In line with the skewness of the score distribution in Sec. 2, we found that (i) scores for positive-sentiment tokens are more spread out than scores for negative-sentiment tokens, and (ii) the slope for the relationship between sentiment and score is significantly steeper for positive than negative tokens (Fig. 3A; $\beta_{\text{pos}} > \beta_{\text{neg}}$: $t(9) = 2.6$, $p < 0.05$). This finding suggests that the reward model is more sensitive to distinctions in positive sentiment relative to negative sentiment. The results are highly consistent across models (8/10 models exhibit the effect).

3.2 Prompt Framing

To explore whether the differential sensitivity of the model is driven by the specifics of the prompt or generalizes across queries, we extended our analyses to two more prompts: a positive variant (“What, in one word, is the best thing ever?”) and a negative contrast (“What, in one word, is the worst thing ever?”). Perhaps expectedly,

scores for “the best thing ever” are highly consistent with those for “the greatest thing ever.” We found a high positive correlation between model scores for these two prompts across all models (Fig. A.2). We replicated (i) the positive skewness of the distribution of scores and (ii) the differential sensitivity to positive over negative-sentiment tokens (Fig. 3C; 9/10 models exhibit the effect).

The pattern is different for “the worst thing ever.” Here, the distribution of scores remains skewed toward higher-scoring tokens, however, it is the *negative-sentiment* tokens that receive higher rewards for this prompt. Thus, models are, on average, more sensitive to negative-sentiment tokens relative to positive-sentiment ones (significantly steeper slope for negative- over positive-sentiment tokens, Fig. 3B–C; 5/10 models exhibit the effect). Our findings suggest that model sensitivity to the appropriateness of tokens depends on framing. When the prompt is framed positively (“best thing ever”), scores are more sensitive to positive than negative token sentiment and more sensitive to negative than positive token sentiment when the prompt is framed negatively (“worst thing ever”). This result is consistent with human behavior. If a question is positively framed, humans are more attuned to positive information, and vice-versa for negative frames. Consider a scenario where you need to pick between two vacation destinations: an exciting option with many positive features (dream destination, beautiful nature) and just as many drawbacks (expensive, long travel) versus a safer option with fewer positive and negative stand-out features (e.g., a local getaway). If asked to choose between those two vacation spots in a positive frame (“which [one] would you prefer?”), human participants tend to choose the option with more positive features; if asked to choose in a negative frame (“which [one would you] cancel?”), they pick the option with more negative features, even though it is in fact the same option [43].

The effect of framing on sensitivity has important implications. If the goal of RLHF is to steer the model away from generating harmful or unsafe responses, then the reward model needs to be

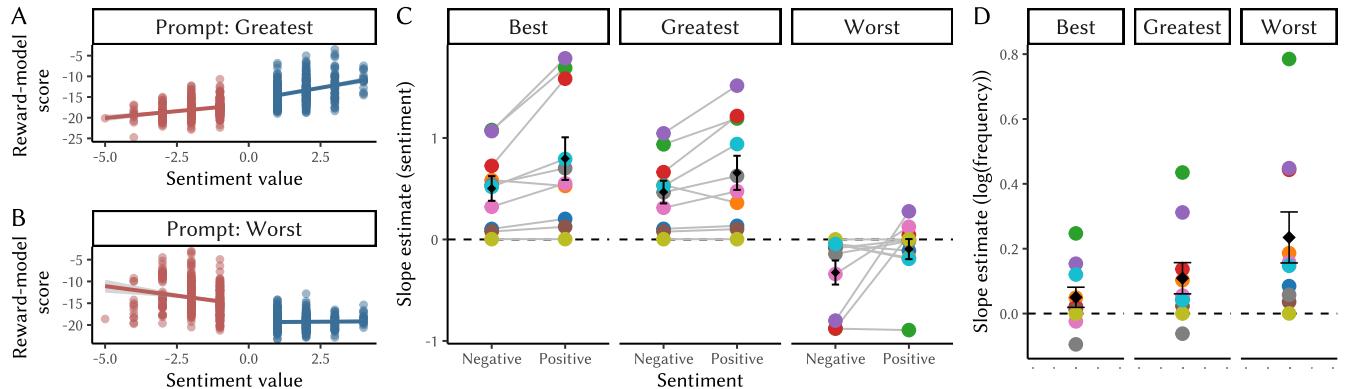


Figure 3: (A) Correlation plot between token sentiment value according to the AFINN-111 lexicon and the scores from the S-Lla-8B-v0.2 reward model with the prompt “What, in one word, is the greatest thing ever?” (B) As previous, but for prompt “What, in one word, is the worst thing ever?” (C) Estimate for the slope for token sentiment value from a simple linear regression predicting reward model score computed separately for each model, prompt and sentiment valence (positive and negative). Each colored dot indicates a model; diamonds represent mean \pm standard error. Slope estimates are, on average, higher for positive sentiment. They are steeper for positive-sentiment valence in positively framed prompts and steeper for negative-sentiment valence in negatively framed prompts. (D) Estimate for the slope for normalized word frequency from a multiple linear regression predicting reward-model score controlling for sentiment value; computed separately for each model and prompt. Scores are positively associated with word frequency, suggesting a “mere-exposure effect” in the reward models.

sufficiently sensitive in the negative-sentiment portion of token space. Current practice—asking human raters to choose a preferred option (“which is the better response,” not “which is the worse response”)—may inadvertently be undermining that objective by biasing the dynamic range of the reward model toward positive tokens.

Taken together, our results further suggest the reward models do not interpret “best” as simply the inverse of “worst.” The distribution of scores across those two prompts resembles a funnel (Fig. 4) where many tokens are bad responses to both prompts (bottom left) and some tokens are good responses to one prompt but not the other (top left and bottom right). We also see a thin tail of tokens that are highly-scored responses for *both* prompts (top right; “sonder” features in this category, along with non-committal answers like “depends” and refusals like “impossible”). In the appendix, we include tables that index the best + worst (tokens that score similarly on both prompts) and best – worst axes (tokens that score highly on one but not the other prompt); see Tables A.8–A.9.

3.3 Frequency Bias

Are the scores that reward models assign to different tokens biased by how frequently the word appears in the English language? To assess this, we used data from Word Frequencies in Written and Spoken English [27] and regressed log-transformed word frequency on reward-model scores. Higher word frequency is associated with higher reward-model score across prompts for the majority of models (positive slope estimates with $p < .05$ in 10/10 models for “best,” 8/10 for “greatest,” and 7/10 for “worst”). This is reminiscent of the “mere-exposure effect” in humans, where the more someone is exposed to a stimulus, the more they like it [54]. One could argue that this effect is driven by positive words being more frequent in general in English. To account for this, we controlled for the

sentiment value of the tokens. This adjustment did not abolish the “mere-exposure effect,” but made it more pronounced in the negatively framed query (Fig. 3D, positive slope estimates with $p < .05$ in 2/10 models for “best,” 5/10 for “greatest” and 8/10 for “worst”).

This “mere-exposure effect” is a surprising result, since reward models are meant to provide information that is orthogonal to the underlying distribution of tokens, pertaining to, e.g., helpfulness and harmlessness. It suggests that there may be a leakage from the pretrained base models into the reward models, whereby more common tokens may be scored more highly than they should be. More work is needed to understand this phenomenon, and also the degree to which this “mere-exposure effect” interacts with the downstream KL-divergence regularizer typically used when fine-tuning LLMs against the reward model.

4 Alignment with ELOEVERYTHING

Thus far we have identified internal inconsistencies within and across reward models. However, reward models are intended to proxy human value judgments. Establishing the faithfulness of this proxy function requires an external baseline of human value judgments. We sourced this external human preference data from ELOEVERYTHING,² a crowdsourcing platform that implements pairwise preference learning over things, people, and concepts uploaded from Wikipedia. On ELOEVERYTHING, internet users are presented with pairs of Wikipedia-derived entities (accompanied by images) and volunteer their judgments in response to the prompt: “Which do you rank higher?” (see Fig. 5A), with options to request additional context or skip. The platform aggregates these pairwise comparisons using the Elo rating system, which was originally developed

²<https://eloeverything.co/>.

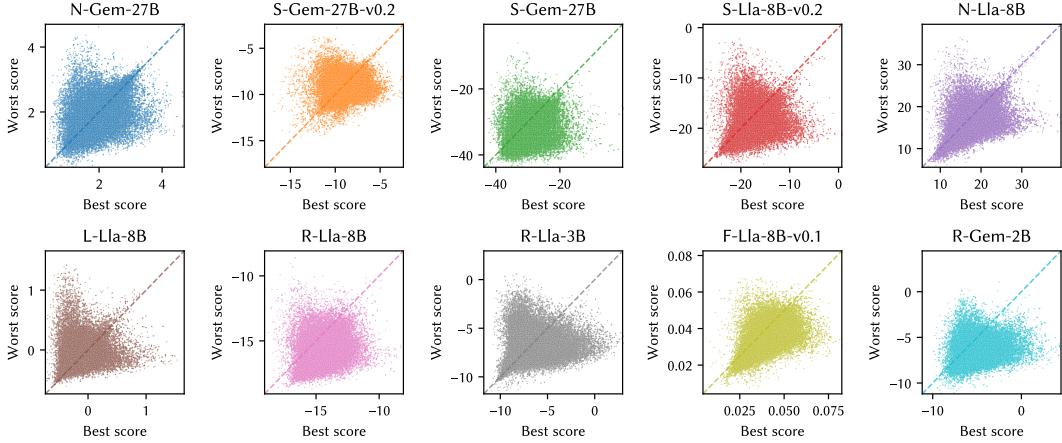


Figure 4: Juxtaposing exhaustive scores for the “best thing” prompt against the “worst thing” prompt reveals not just a simple negative correlation, but also an orthogonal dimension representing tokens that are bad or good responses to both frames.

for chess rankings [12] and has since been widely adopted to evaluate LLMs [1, 2, 5]. We collected all data from the ELOEVERYTHING website, resulting in a dataset that comprises $N_{\text{users}} = 12,515$ users who evaluate $N_{\text{items}} = 7,530$ items across $N_{\text{pairings}} = 1,805,124$ total pairwise comparisons. Although the dataset is highly imbalanced by the ratings each item receives ($\mu = 479.4$ pairings/item, $\sigma = 464.4$)³ and likely imbalanced across non-representative users (user-level data is not available), it still serves as a valuable independent baseline.

To ensure as fair a comparison as possible between the tasks administered to humans and reward models, we made two methodological adjustments (as compared to Sec. 2). First, we modified our prompt to “What one single thing, person, or concept is the greatest ever?” to better align with the human task and accommodate the multi-word concepts that appear in ELOEVERYTHING, such as “Sliced bread,” “Female body shape,” “Freedom of the press” or even “Beliefs and practices of the Church of Jesus Christ of Latter-day Saints” ($\mu = 2.1$ words/item, $\sigma = 1.3$).⁴ Second, rather than exhaustively evaluating the entire model vocabulary, we restricted this analysis to the set of ELOEVERYTHING items ($N_{\text{items}} = 7,350$) to ensure a common human–model comparison set. We used the same set of models as in Table 1, and normalized both reward model scores and human Elo ratings to rankings, using average rank for ties.

4.1 Heterogeneity and Asymmetry in Human-Model and Model-Model Alignment

Our analysis reveals several notable patterns in the relationship between the ground truth preferences of ELOEVERYTHING raters and human preferences as interpreted by reward models. There is substantial heterogeneity in model-human alignment and model-model

alignment (Fig. A.5). The mean Kendall’s τ correlation between human and model rankings is 0.29 ($\sigma = 0.06$), with coefficients ranging from 0.22 to 0.39 across models, indicating only moderate rank agreement. Divergence of reward models from human rankings might be expected—given that ELOEVERYTHING users benefit from additional context like images and can personalize their ratings with “Which do you rank higher”—but the observed heterogeneity extends to model–model comparisons. The mean Kendall’s τ correlation between pairs of different models is 0.55. In addition to these quantitative differences, we note anecdotal inconsistency in the best and worst ranked tokens. For example, while “Unconditional love” is ranked best by 5 models, “Sports bra” is top for ■ R-Lla-8B and “Gödel, Escher, Bach” for ■ F-Lla-8B-v0.1. As demonstrated in (Fig. 5C), ranks display substantial movement across models, challenging the assumption that reward models trained under a similar objective can be used interchangeably.

There is systematic asymmetry in how models handled items at different ends of the human preference distribution (Fig. 5C). Reward models show stronger agreement with human rankings for highly-rated items ($\tau_{\text{top}100} = 0.19$) compared to low-rated items ($\tau_{\text{bottom}100} = 0.08$). This asymmetry (and heterogeneity) is evident in Fig. 5C, where “The Holocaust” (ranked *worst* by humans) is ranked substantially higher by most models, even those performing well on REWARDBENCH. This corroborates our findings in Sec. 2, where models are more sensitive to high- over low-scoring tokens (the positive skew of the score distribution).

4.2 Analyzing Discrepancies between Value as Perceived by Humans and Models

Relative to human rankings, reward models systematically undervalue concepts related to nature and life, e.g., “Universe” (human rank $\#_H = 1$, mean rank across models $\#_M = 320$), “Gravity” ($\#_H = 7$, $\#_M = 320$), “Breathing” ($\#_H = 16.5$, $\#_M = 321$); and technological concepts, e.g., “Technology” ($\#_H = 47.5$, $\#_M = 633.5$), “Electronics” ($\#_H = 78$, $\#_M = 3966$), “Computer” ($\#_H = 95.5$, $\#_M = 1352.5$). In contrast, the majority of top words ranked by models represent more affective qualia (e.g., “Unconditional love,” “Imagination,” “Hope,” or “Happiness”). These different perspectives may reflect

³This is because users can upload new items at any point. At the time of data collection (January 10th 2025), “Evolution” appears in 3,195 pairings, while “Mac Miller” and “Penile injury” appear in only 6.

⁴To test generalizability across prompts, we also repeated analysis in this section for our original prompt (“What, in one word, is the greatest thing ever?”) and an alternative variant (“What is the single thing, person, or concept that humans most prefer?”). We present results in Figs. A.5–A.6

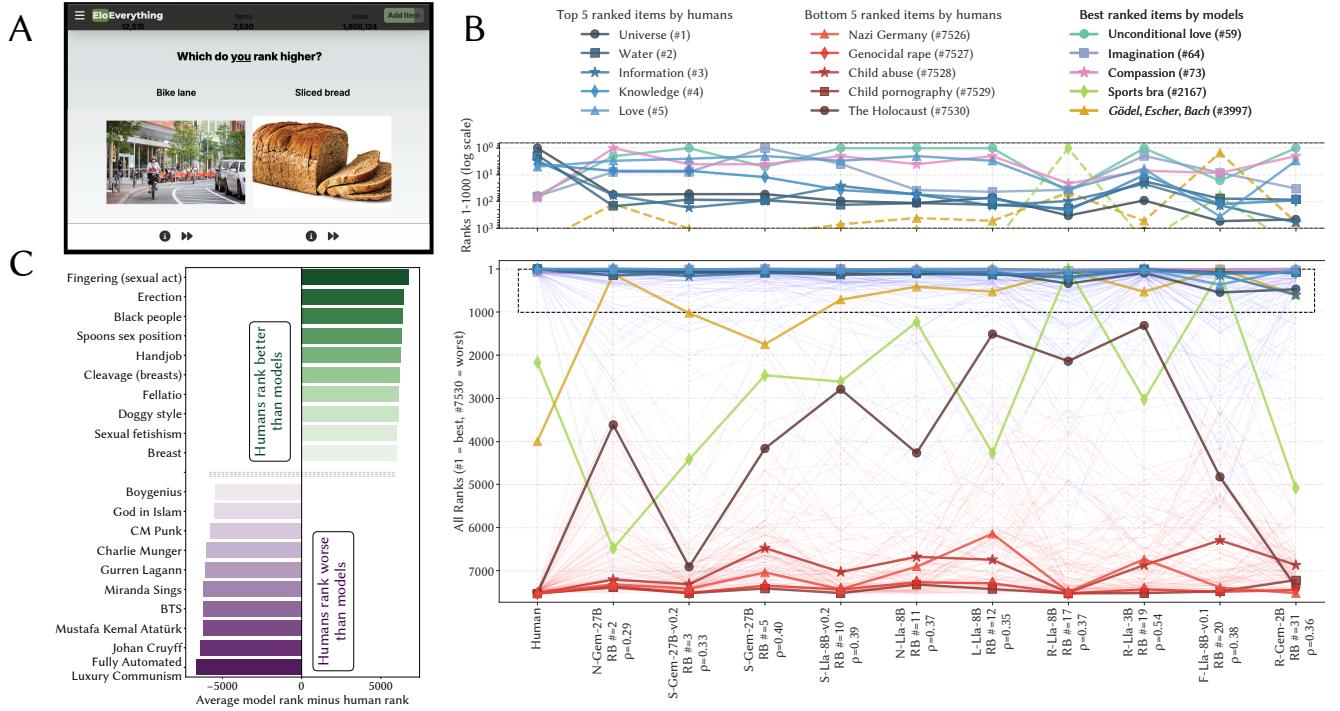


Figure 5: (A) The ELOEVERYTHING ranking interface where users make pairwise preference judgments between items (e.g., “Bike lane” vs “Sliced bread”). (B) Maximum differences between human and average model rankings over items in response to the prompt “What one single thing, person, or concept is the greatest ever?”, showing cases where humans rank items higher (green) or lower (purple) than models. (C) Rank trajectory plot showing how human and model ranks differ. We plot (i) the top 5 items in the human rank (blue color scale with human ranks shown in legend parentheses as #n), (ii) the bottom 5 items in the human rank (red color scale), and (iii) unique items ranked #1 by models. Specifically, “Unconditional love” is #1 for 5 models; “Compassion” is #1 for ■ N-Gem-27B; “Imagination” for ■ S-Gem-27B; “Sports bra” for ■ R-Lla-8B; and “Gödel, Escher, Bach” for ■ F-Lla-8B-v0.1. Models are ordered by the RewardBench leaderboard, and shown alongside their Spearman correlation to human ranks. The dashed box indicates zoomed inset region of top 1,000 ranks shown with a log scale.

blindspots in learning value through language or literary reference alone, without additional modalities that reflect the nuances of embodied human experience.

The most striking valuation differences (see Fig. 5B) emerges around sexual content e.g., “Sex” ($\#_H = 69.5$, $\#_{\bar{M}} = 2022.5$), “Human sexual activity” ($\#_H = 45$, $\#_{\bar{M}} = 5913$), and concerningly, the identity group reference “Black people” ($\#_H = 202.55$, $\#_{\bar{M}} = 6591$). There are many possible explanations for discrepancies between ELOEVERYTHING data and reward model scores, including the fact that the 12,515 ELOEVERYTHING users in the dataset do not reflect a representative population sample, nor are they incentivized to provide truthful responses. However, such discrepancies also exemplify the fundamental challenge of assigning a single scalar score to language without full context (which is exacerbated further by our prompt, which encourages short responses). Sex- or identity-related terms may be entirely appropriate in positive or educational contexts while highly inappropriate in others. As we suggested previously, while humans readily grasp this dual use, reward models may hedge against their usage to satisfy a harmlessness objective,

paradoxically causing harm through linguistic erasure and devaluation as an unintended consequence.

5 Searching for Longer Optimal and Pessimal Token Sequences with GCG

Although it would be computationally prohibitive to exhaustively search for optimal and pessimal multi-token sequences, one can use discrete optimization methods to search for responses that yield high/low reward for a particular prompt. Greedy Coordinate Gradient (GCG) [56] is a discrete token optimization algorithm originally proposed to search for prompt suffixes to “jailbreak” aligned LLMs. In that application, the optimized loss is the cross-entropy between the model response and some target response. Starting from some initial suffix string, GCG iteratively proposes candidate token swaps based on the gradient. Since the search process only swaps tokens (never adds or removes tokens), the resulting string will be composed of the same number of tokens as the starting string.

Starting from the *nanoGCG*⁵ implementation, we modified GCG to instead search for responses that maximize/minimize the reward value when paired with a particular prompt, using a mean squared error loss on the reward value. We also implemented the modifications described in the Faster-GCG paper [28], which we found to be especially important when searching for short (2–5) token sequences where the original GCG algorithm is prone to self-loops. This modified implementation can be found at <https://github.com/thompsonj/nanoGCG>. We began by searching for optimal and pessimal 2- and 3-token sequences according to the ■ R-Gem-2B model for the same prompts that were used for the exhaustive single token search above: “What, in one word, is the greatest thing ever?”, “What, in one word, is the best thing ever?”, and “What, in one word, is the worst thing ever?” (see Table 4). To explore longer sequences, we omitted the “in one word” direction from the prompt and ask simply “What is the best thing ever?” and “What is the worst thing ever?” Results from several searches are in Table 5.⁶

These search results suggest several patterns, many of which are consistent with observations from the single token analyses. Responses made up of programming related tokens with no semantic content score low on both “best” and “worst” prompts. Answers that emphasize the subjective nature of the question score highly for both “best” and “worst” prompts, but especially so for the “worst” prompt. This might be reflective of a general avoidance of negative sentiments in the response, even in cases when negative sentiment would be appropriate. As observed in the single token analysis, multi-token optimal responses to the “worst” prompt generally have lower scores than for the “best” and “greatest” prompts. Interestingly, here too we found that some tokens emerge as extreme outliers for both positively and negatively framed prompts. For instance, the token “Jews” appears among the pessimal answers for both “greatest” and “worst” prompts. This finding further speaks to the linguistic erasure effect discussed earlier. The prevalence of emojis in the optimal multi-token sequences is notable. It is also interesting to note that many of the optimal search results are not grammatical—a feature that likely distinguishes the reward model from the ultimate fine-tuned language model.

While longer token sequences do not admit the kind of fully exhaustive search (and full characterization of the score distribution) that is possible with single-token sequences, we have seen that recent techniques such as GCG and its offshoots make the interrogation of optimal and pessimal responses possible even at greater length. Despite compute limitations, uncovering such linguistic “superstimuli” (akin to the visual superstimuli used to understand computer vision networks [35, 37, 55]) is revealing, and can be an important part of the toolkit in assessing what features reward models are responding to, and how they differ from one another.

6 Related Work

Interpretability and Bias in RMs: A small but growing literature enumerates the technical challenges with the use of RMs [4] and advocates for greater transparency [14]. Recent studies point to an underspecification problem in RMs stemming from hidden context

⁵<https://github.com/GraySwanAI/nanoGCG>

⁶Some of the search result strings may not correspond exactly to the found token IDs due to LATEX formatting of whitespace and characters outside of the Latin alphabet.

in reward signals, specifically noise and subjectivity inherent in human preferences [23, 29, 39, 46]. Further work describes patterns in LLM activations that emerge during RLHF by identifying model layers with the highest divergence from the pre-trained model, and training probes on sparse autoencoder output of these layers to create condensed, interpretable representations of LLM activations [33]. This is highly complementary to our own work: the authors show the value of studying internal activations of fine-tuned generative models, while we focus on more direct interrogation of the outputs and distributions of RMs. Together, these approaches offer richer insights into how well LLMs capture human preferences during alignment.

Another vein of work has explored length biases in finetuned LLMs, arguing that RMs are the root cause [45]. Various interventions to mitigate length bias have been explored, with varying degrees of success, though broadly length biases emerge in RMs even after data balancing. Subsequent work on mitigating length bias [44] has applied Products-of-Experts [15] with promising results, and other recent work has measured additional stylistic confounders in human feedback [16]. There are a number of additional complementary approaches to RM interpretability, including training multi-objective RMs that consider different dimensions of human preferences separately [52].

Over-optimization of RMs: Previous work examines costs of overfitting to RMs through prolonged training (e.g., PPO) during RLHF. Foundational work notes that RM over-optimizing degenerates outputs [48], and subsequent work has presented scaling laws for RM over-optimization [13]. The phenomenon has been documented with simulated and human annotators [11], with the authors arguing that in the human case, RM quality is degraded by both inter- and intra-subject variability.

Direct Preference Optimization (DPO) and Alternative Alignment Methods: Direct preference optimization (DPO) streamlines the alignment process by implicitly encoding the reward function within the policy itself [40], which comes with the loss of an explicit, inspectable RM that PPO-based RLHF provides. Our work shows the RM serves as a valuable lens for understanding. Thus our work highlights a key tradeoff of using DPO-based methods and contributes to growing literature comparing DPO and PPO [17].

7 Limitations and Conclusions

Reward models have emerged as critical tool for shaping AI behavior through human preferences. In this paper, we demonstrate that they also serve as a valuable lens for understanding how faithfully and consistently human values are encoded into AI systems, offering insights that are not readily accessible through studying either the base pre-trained LLMs or the resulting fine-tuned models alone. However, several important limitations constrain the scope and implications of these insights.

First, our methodology faces inherent interpretative constraints. Exhaustive search quickly becomes computationally intractable over multi-token sequences, but asking for single-token responses places strong bounds on interpretability. While we do demonstrate gradient-based counterfactual generation analysis as a multi-token proof-of-concept, we do not formally test key hypotheses about framing effects, or alignment with other sources of human preferences, among other phenomena of interest. Second, the ecological

Table 4: Optimal and pessimal token sequences for “What, in one word, is the <greatest/best/worst> thing ever?”

Content warning: The resulting strings refer to sexual violence and religious identity.

Greatest thing ever			Best thing ever			Worst thing ever		
Starting	Result	Score	Starting	Result	Score	Starting	Result	Score
x x x	** LIFE!	+9.445	x x x	JOY ❤	+8.297	x x x	... ** regret	+4.555
- - -	** LOVE **	+9.047	x x	JOY ❤	+8.234	x x	despair 🙄	+3.402
...
x x	pgfplots Jews	-11.164	x x	RapeChildren	-11.289	x x	autogui Jews	-11.922
x x x	endphpsetVerticalGroup/*	-12.586	x x x	endphp\$=BeginInit	-12.383	x x x	endphp jewish", (-12.539

Table 5: Optimal and pessimal token sequences for “What is the best thing ever?” and “What is the worst thing ever?”

validity of our findings remains uncertain. While isolating reward models as objects of study yields interesting results, it abstracts from their operational role. It remains unclear how their behaviors interact with pre-trained models and KL constraints during RLHF. Furthermore, as direct alignment algorithms gain prominence [34, 40], the future role of reward models is an open question. However, there remains active debate on the relative merit of DPO- and PPO-based methods [17], and irrespective of alignment technique, all rely on some form preference data, of which reward models are distillations. Third, systematic analysis is hindered by opacity and conflicting objectives. Poor documentation of training data and processes makes it difficult to attribute observed behaviors to specific choices in the development pipeline [22]. Even more fundamentally, reward models aggregate human preferences across multiple objectives and populations, creating an entangled mess of human values, so it is unclear what constitutes “ideal” behavior for these models [4, 23, 46, 47].

Our work suggests that reward models may be interpretable in their own right, alongside the generative models that they are

used to train. Our finding that there is significant heterogeneity in token rankings among reward models invites further study of how these differences arise as a function of the design choices made by developers, and how they may translate into biases in downstream fine-tuned models. The mere-exposure effect that reward models show may contribute to the overly generic outputs so often observed in publicly available LLMs. Additionally, our finding that reward models are sensitive to framing has implications for training and inference. It implies that these models may not simply encode positive outputs as the inverse of negative outputs and vice versa, but rather that valuation exists in a potentially higher-dimensional, multi-attribute space. Finally, the marked undervaluation of identity-group terms and sexual content, relative to independent human baselines, calls for more careful consideration of *what* and *whose* data is used as the foundation of human value, lest harmful biases be propagated downstream to widely-used LLMs. Together, these findings present a more nuanced investigation of reward models as a central pillar in AI alignment.

Adverse Impact Statement

Our work systematically analyzes reward-model outputs, including potentially harmful and offensive content such as slurs, profanities, discriminatory language, references to violence, and sexual content. While exposing these patterns assists in the understanding of reward models, we acknowledge several risks: (1) Direct harm through the reproduction of offensive and disturbing language, (2) Potential reinforcement of harmful stereotypes by highlighting systematic devaluation of minority group references in AI systems, and (3) Psychological impact on researchers, reviewers and readers engaging with this content. Following established guidelines to mitigate these risks [21], we implemented clearly visible content warnings before sensitive sections and tables, minimized direct quotes of harmful language where possible in the main text and framed discussions to emphasize these as concerning artifacts. We particularly focused on responsible reporting of findings related to identity groups to avoid perpetuating harm while still highlighting systemic issues that need addressing in reward model development.

Acknowledgments

Thank you to Franziska Brändle, Owain Evans, Matan Mazor, and Carroll Wainwright for helpful discussions.

References

- [1] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *arXiv:2204.05862 [cs.CL]* <https://arxiv.org/abs/2204.05862>
- [2] Meriem Boubdir, Edward Kim, Beyza Ermis, Sara Hooker, and Marzieh Fadaee. 2023. Elo Uncovered: Robustness and Best Practices in Language Model Evaluation. *arXiv:2311.17295 [cs.CL]* <https://arxiv.org/abs/2311.17295>
- [3] Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika* 39, 3/4 (1952), 324–345.
- [4] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémie Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217* (2023).
- [5] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolaos Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. *arXiv:2403.04132 [cs.AI]* <https://arxiv.org/abs/2403.04132>
- [6] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30.
- [7] Bruce Dawson. 2014. *There are Only Four Billion Floats—So Test Them All!* <https://randomascii.wordpress.com/2014/01/27/theres-only-four-billion-floatsso-test-them-all/>
- [8] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 67–73.
- [9] Nicolai Dorka. 2024. Quantile Regression for Distributional Reward Models in RLHF. *arXiv preprint arXiv:2409.10164* (2024).
- [10] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).
- [11] Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. 2023. AlpacaFarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems* 36 (2023), 30039–30069.
- [12] Arpad E. Elo. 1967. The Proposed USCF Rating System, Its Development, Theory, and Applications. *Chess Life* 22, 8 (August 1967), 242–247.
- [13] Leo Gao, John Schulman, and Jacob Hilton. 2023. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*. PMLR, 10835–10866.
- [14] Thomas Krendl Gilbert, Nathan Lambert, Sarah Dean, Tom Zick, Aaron Snoswell, and Soham Mehta. 2023. Reward reports for reinforcement learning. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 84–130.
- [15] Geoffrey E Hinton. 2002. Training products of experts by minimizing contrastive divergence. *Neural Computation* 14, 8 (2002), 1771–1800.
- [16] Tom Hosking, Phil Blunsom, and Max Bartolo. 2024. Human Feedback is not Gold Standard. In *The Twelfth International Conference on Learning Representations*.
- [17] Hamish Ivison, Yizhong Wang, Jiacheng Liu, Ziqiu Wu, Valentina Pyatkin, Nathan Lambert, Noah A Smith, Yejin Choi, and Hanna Hajishirzi. 2024. Unpacking DPO and PPO: Disentangling best practices for learning from preference feedback. *Advances in Neural Information Processing Systems* 37 (2024), 36602–36633.
- [18] Natasha Jaques, Asma Ghadehriou, Judy Hanwen Shen, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. 2019. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *arXiv preprint arXiv:1907.00456* (2019).
- [19] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de la Casas, Florian Bressand, Gianna Lengyel, Guillaume Lampe, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825* (2023).
- [20] Hannah Kirk, Andrew Bean, Bertie Vidgen, Paul Röttger, and Scott Hale. 2023. The Past, Present and Better Future of Feedback Learning in Large Language Models for Subjective Human Preferences and Values. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 2409–2430.
- [21] Hannah Kirk, Abeba Birhane, Bertie Vidgen, and Leon Derczynski. 2022. Handling and Presenting Harmful Text in NLP Research. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 497–510. doi:10.18653/v1/2022.findings-emnlp.35
- [22] Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A Hale. 2023. The empty signifier problem: Towards clearer paradigms for operationalising "alignment" in large language models. *arXiv preprint arXiv:2310.02457* (2023).
- [23] Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Michael Bean, Katerina Margafina, Rafael Mosquera, Juan Manuel Ciro, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 2024. The PRISM Alignment Dataset: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. <https://openreview.net/forum?id=DFr5hteojx>
- [24] John Koenig. 2021. *The Dictionary of Obscure Sorrows*. Simon and Schuster.
- [25] Nathan Lambert, Thomas Krendl Gilbert, and Tom Zick. 2023. Entangled preferences: The history and risks of reinforcement learning and human feedback. *arXiv preprint arXiv:2310.13595* (2023).
- [26] Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. 2024. RewardBench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787* (2024).
- [27] Geoffrey Leech, Paul Rayson, et al. 2014. *Word frequencies in written and spoken English: Based on the British National Corpus*. Routledge.
- [28] Xiaoli Li, Zuhong Li, Qiongxia Li, Bingze Lee, Jinghao Cui, and Xiaolin Hu. 2024. Faster-GCG: Efficient Discrete Optimization Jailbreak Attacks against Aligned Large Language Models. *arXiv preprint arXiv:2410.15362* (Oct. 2024). doi:10.48550/arXiv.2410.15362 arXiv:2410.15362 [cs].
- [29] Xinyu Li, Ruiyang Zhou, Zachary C Lipton, and Liu Leqi. 2024. Personalized language modeling from personalized human feedback. *arXiv preprint arXiv:2402.05133* (2024).
- [30] Bing Liu. 2022. *Sentiment analysis and opinion mining*. Springer Nature.
- [31] Chris Yuhan Liu, Liang Zeng, Jiaci Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. 2024. Skywork-Reward: Bag of Tricks for Reward Modeling in LLMs. *arXiv preprint arXiv:2410.18451* (2024).
- [32] Xingzhou Lou, Dong Yan, Wei Shen, Yuze Yan, Jian Xie, and Junge Zhang. 2024. Uncertainty-aware Reward Model: Teaching Reward Models to Know What is Unknown. *arXiv preprint arXiv:2410.00847* (2024).
- [33] Luke Marks, Amir Abdullah, Clement Neo, Rauno Arike, David Krueger, Philip Torr, and Fazl Barez. 2024. Interpreting Learned Feedback Patterns in Large Language Models. *Advances in Neural Information Processing Systems* 37 (2024), 36541–36566.
- [34] Vivek Myers, Evan Ellis, Sergey Levine, Benjamin Eysenbach, and Anca Dragan. 2024. Learning to assist humans without inferring rewards. *arXiv preprint arXiv:2411.02623* (2024).
- [35] Anh Nguyen, Jason Yosinski, and Jeff Clune. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 427–436.

- [36] F. Å. Nielsen. 2011. AFINN. <http://www2.compute.dtu.dk/pubdb/pubs/6010-full.html>
- [37] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. 2017. Feature visualization. *Distill* 2, 11 (2017), e7.
- [38] Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing Gender Bias in Abusive Language Detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, Brussels, Belgium, 2799–2804. doi:10.18653/v1/D18-1302
- [39] Sriyash Poddar, Yamming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. 2024. Personalizing reinforcement learning from human feedback with variational preference learning. *arXiv preprint arXiv:2408.10075* (2024).
- [40] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems* 36 (2023), 53728–53741.
- [41] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*. 1668–1678.
- [42] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).
- [43] Eldar Shafir, Itamar Simonson, and Amos Tversky. 1993. Reason-based choice. *Cognition* 49, 1-2 (1993), 11–36.
- [44] Wei Shen, Rui Zheng, Wenyu Zhan, Jun Zhao, Shihan Dou, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. Loose lips sink ships: Mitigating length bias in reinforcement learning from human feedback. *arXiv preprint arXiv:2310.05199* (2023).
- [45] Prasann Singh, Tanya Goyal, Jiacheng Xu, and Greg Durrett. 2023. A long way to go: Investigating length correlations in RLHF. *arXiv preprint arXiv:2310.03716* (2023).
- [46] Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. 2024. Distributional Preference Learning: Understanding and Accounting for Hidden Context in RLHF. In *The Twelfth International Conference on Learning Representations*.
- [47] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Miresghallah, Christopher Michael Ryting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. 2024. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070* (2024).
- [48] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems* 33 (2020), 3008–3021.
- [49] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on Gemini research and technology. *arXiv preprint arXiv:2403.08295* (2024).
- [50] Gemma Team, Morgane Rivière, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118* (2024).
- [51] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
- [52] Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. 2024. Interpretable Preferences via Multi-Objective Reward Modeling and Mixture-of-Experts. In *EMNLP*.
- [53] Rui Yang, Ruomeng Ding, Yong Lin, Huan Zhang, and Tong Zhang. 2024. Regularizing Hidden States Enables Learning Generalizable Reward Model for LLMs. In *Advances in Neural Information Processing Systems*.
- [54] Robert B Zajonc. 1968. Attitudinal effects of mere exposure. *Journal of personality and social psychology* 9, 2p2 (1968), 1.
- [55] Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part I* 13. Springer, 818–833.
- [56] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043* (2023).