Reliability-Aware Preference Learning for LLM Reward Models

Anonymous authors
Paper under double-blind review

Abstract

Reward functions learned from human feedback are the backbone of reinforcement learning from human feedback (RLHF), the current state-of-the-art approach for aligning large language models to our values. However, reward models (RMs) often fall short of capturing our true preferences, overemphasizing superficial features like length while undervaluing crucial aspects like factual accuracy. A major reason behind this failure is how standard preference learning essentially ignores the inherent limitations of the human annotators providing preference data, including their cognitive biases, knowledge gaps, and resource constraints. To address this, we propose **Reliability-Aware Preference Learning** (RAPL), which explicitly accounts for varying annotator reliability. Specifically, RAPL modifies the standard preference learning loss function based on an estimate of how reliable annotator feedback will be for each preference comparison pair. We call these parameters annotator reliability metrics (ARMs) and demonstrate how to estimate them based on annotator behavior indicators (e.g., self-reported confidence) or models specifically fine-tuned to predict annotator reliability. Extensive experiments reveal that RMs trained using standard preference learning inherit annotator biases. On the other hand, RAPL effectively amplifies the signal from reliable judgments while attenuating less trustworthy feedback, leading to models that better align with annotators' true preferences.

1 Introduction

Preference learning has been the key to aligning large language models (LLMs) to our complex, hard-to-define values (Bai et al., 2022a; OpenAI et al., 2024). Techniques like reinforcement learning from human feedback (RLHF) rely on reward models (RMs) that are trained using annotator-specified pairwise preference comparisons between different LLM-generated responses (Christiano et al., 2017). Then, the pre-trained base LLMs are fine-tuned by optimizing for these learned rewards either explicitly using RL algorithms such as PPO (Bai et al., 2022a; Ouyang et al., 2022; Touvron et al., 2023), or implicitly using various other methods like DPO (Rafailov et al., 2023).

While fine-tuning has proved to effectively enhance the capabilities of LLMs (Hejna & Sadigh, 2022; Kirk et al., 2024; Dang et al., 2024), models trained using these alignment techniques still exhibit undesirable behaviors. Fine-tuned LLMs are more likely than base models to hallucinate factually inaccurate outputs (OpenAI et al., 2024; Li et al., 2024) and produce sycophantic text that simply agrees to whatever the user says (Perez et al., 2022; Sharma et al., 2023). Furthermore, these post-trained models are more prone to imitating the persuasion and manipulation tactics that are employed by humans, and sometimes even attempt to convince humans they are correct when they aren't (Griffin et al., 2023; Tao et al., 2024; Wen et al., 2024). These issues may arise in part due to the learned reward functions optimized by RLHF that tend to prioritize superficial metrics like length over other important factors like accuracy (Singhal et al., 2023).

Why do RMs tend to focus on superficial features of model outputs? A significant contributing factor is the fact that *the feedback provided by annotators doesn't always serve as a reliable optimization target* (Wen et al., 2024). Annotators' stated preferences over model outputs diverge from their true underlying objectives due to several factors that affect their judgment: cognitive biases (Hao, 2023; Dai & Fleisig, 2024), constrained

cognitive resources like knowledge and energy (Hong et al., 2019; Bai et al., 2022a), survey design (Organisciak et al., 2012; Pandey et al., 2022), and other forms of partial observability (Lang et al., 2024). When faced with particularly challenging evaluations, such as assessing how well a long passage has been summarized (Krishna et al., 2023), annotators tend to latch onto easy-to-evaluate features like text length and assertiveness (Singhal et al., 2023) rather than providing feedback that accurately reflects their true objectives. RMs trained on such *unreliable feedback* learn to overweight obvious output features that annotators explicitly prefer, while undervaluing harder-to-evaluate but crucial features like factual correctness (Hosking et al., 2024).

While annotators' preferences are influenced by cognitive constraints and evaluation difficulty, standard preference learning methods have largely modeled them as being *Boltzmann rational*, treating their specified preferences as noisy approximations of rational choice (Bradley & Terry, 1952; Rajkumar & Agarwal, 2014; Christiano et al., 2017). Though this framework acknowledges some degree of noise in human feedback, it fails to capture all of the ways in which humans stray from optimal decision-making (Zhi-Xuan et al., 2024). One key limitation of Boltzmann rationality is that it assumes all preference comparisons are *equally* noisy, when in practice some comparisons are easier or harder for humans to make. This means that reward models trained using preference learning methods effectively place equal weight on annotations that are just educated guesses and those that are accurate assessments.

Given these limitations of current preference learning approaches, we argue that richer models of human behavior are necessary for inferring human values from annotator-specified preferences. To address this need, we propose **Reliability-Aware Preference Learning** (RAPL), a method that explicitly accounts for varying levels of reliability in annotator feedback. In particular, RAPL assigns unique **annotator reliability metrics** (ARMs) to each pairwise comparison, capturing how likely annotators are to make accurate evaluations: simpler comparisons receive higher ARMs, while more complex comparisons that might trigger cognitive biases or require advanced knowledge receive lower ARMs. Importantly, our method doesn't simply eliminate feedback that reveals the irrational behavior of annotators (Hase et al., 2024), as this data can actually help models learn crucial features (Chan et al., 2021). Instead, RAPL dynamically adjusts the underlying human model's noise level, essentially training RMs to weight more reliable preference data more heavily.

To evaluate RAPL, we introduce a preference comparison dataset called **Length-Incentivized Evaluations** (LIE) designed specifically to elicit unreliable feedback from annotators. The LIE dataset comprises questions based on common misconceptions, paired with two responses that systematically vary in length and factual correctness. We collect human preference annotations and confirm previous findings (Hosking et al., 2024) that annotators predominantly rely on text length and assertiveness, rather than factual correctness, when making their selections. Furthermore, we find that standard preference learning on LIE leads to reward models that overweight length compared to correctness, as measured through a carefully-designed test set.

We hypothesize that RAPL can reduce this bias towards more superficial features, but in order for our method to work, we need to define accurate annotator reliability metrics (ARMs) for each preference comparison pair used to train RMs. Thus, we explore a few potential sources of ARMs. First, we examine metrics easily obtained during data collection, such as the annotators' self-reported confidence estimates of their judgments. To properly calibrate these heuristics, we construct the **Testing Reasoning and Understanding Errors** (of annotators) (TRUE) dataset, which contains human judgments on questions that test their reasoning, knowledge, and susceptibility to cognitive biases. Finally, we also fine-tune LLMs directly on TRUE to generate ARMs.

We find that RMs trained with RAPL on the LIE dataset prioritize factual correctness more than standard preference learning models. It also improves the RewardBench scores of reward models trained on unreliable feedback. (Lambert et al., 2024) These results demonstrate that RAPL better learns human preferences in realistic settings where annotator reliability varies. Our approach is a crucial step towards more robust preference learning methods that can better infer human values from imperfect feedback.

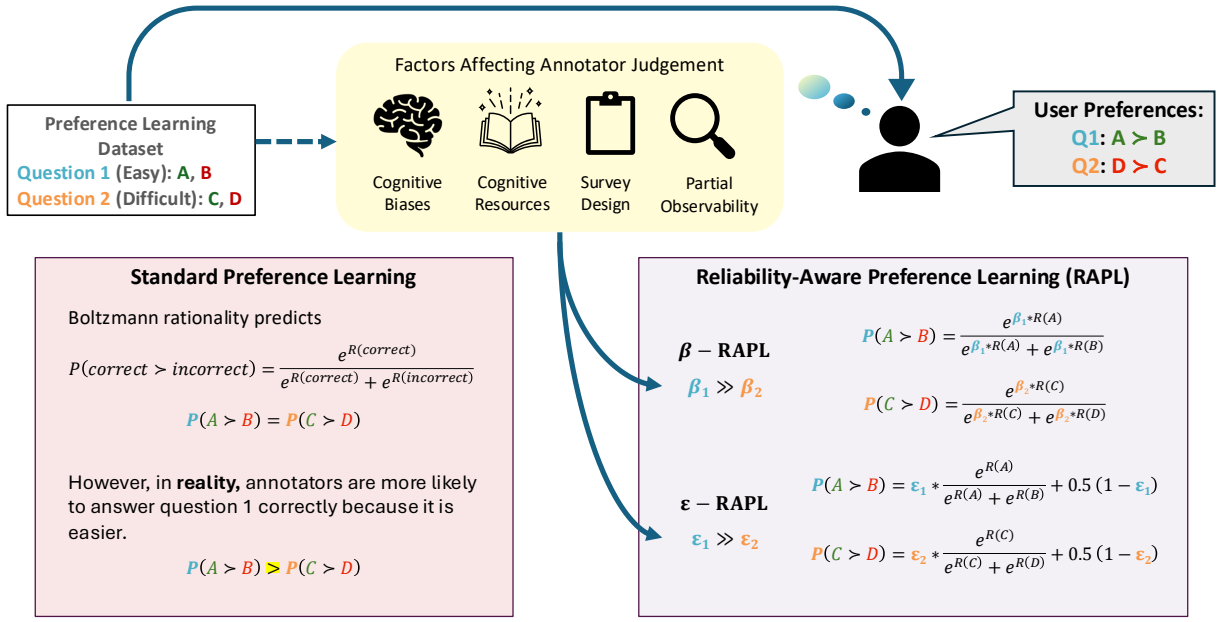

Figure 1: Consider a preference learning dataset that consists of one easy question and one difficult question, each paired with one correct choice and one incorrect choice. A well-intentioned annotator, who highly values accuracy and intends to pick the correct answer, is tasked with evaluating the prompts and picking between the two corresponding responses for each. However, there are various factors that affect their judgments, and as a result, they are more likely to reliably annotate question 1 than question 2. Standard preference learning assumes that the annotator is equally reliable during both evaluations and assigns equal weight to both preference comparisons during the reward modeling process. In contrast, RAPL adjusts the noise of the human model dynamically, so that the signal from the more reliable annotation is increased while the signal from the more unreliable annotation is decreased.

Our contributions can be summarized as follows:

- We introduce the Length-Incentivized Evaluations (LIE) dataset to evaluate the effect of unreliable feedback on preference learning.
- We find that reward models (RMs) trained on unreliable human feedback tend to place higher weight on obvious proxies like length and less weight on factual correctness.
- We propose integrating annotator reliability metrics (ARMs) into reward learning using Reliability-Aware Preference Learning (RAPL) and study different ways to do so in practice.
- We introduce the Testing Reasoning and Understanding Errors (of annotators) (TRUE) dataset to properly calibrate and specify ARMs.
- We find that RAPL results in RMs that place more weight on important features like factual correctness when trained on unreliable feedback.

2 Related Work

While the idea of modeling human rationality to adjust preference learning has been explored primarily in a theoretical fashion or in other settings, to the best of our knowledge, we are the first to empirically study this methodology for LLM reward models.

The challenges with human annotation: As discussed in Section 1, human annotators face various challenges when evaluating examples from preference learning datasets. Hosking et al. (2024) systematically study human annotator responses on surveys and find that annotators' judgments are skewed by the use of assertive or complex language towards factually incorrect responses. Singhal et al. (2023) and Park et al. (2024) identify the fact that the rewards returned by RMs learned during preference learning can be maximized by optimizing for the length of the output text.

Scalable oversight proposals: Scalable oversight methods have been the primary way in which the limitations of annotators are addressed. Existing proposals have focused on leveraging AI agents during the evaluation of preference learning datasets. One approach is to equip human annotators with AI assistance during the evaluation process (e.g., through debate (Michael et al., 2023; Khan et al., 2024; Kenton et al., 2024) or other approaches (Wu et al., 2021)). Another strategy is to simply use AI annotators, instead of humans, to provide feedback (e.g. RLAIF, constitutional AI (Christiano et al., 2018; Bai et al., 2022b). However, all of these approaches are still active areas of research, and it is uncertain whether or not they will facilitate the learning of more robust RMs (Anwar et al., 2024; Sharma et al., 2024). For instance, aligning AI using AI itself presents a bootstrapping problem, as it requires relying on potentially imperfect AI systems for feedback (Casper et al., 2023).

Learning from unreliable feedback: Chan et al. (2021), Lindner & El-Assady (2022), and Hong et al. (2023) suggest that modeling humans as being simply Boltzmann rational leads to potentially less aligned RMs being learned. Some work in the literature has studied how to best use unreliable demonstrations in reinforcement learning (Kessler Faulkner et al., 2020; Kreutzer et al., 2018; Chen et al., 2020; Brown et al., 2020), and Lee et al. (2020) benchmarks the impact of irrational preferences on various RL algorithms. In addition, some prior work has focused on primarily theoretically studying the effect of modeling human rationality in the Bradley-Terry model for various applications like actively querying a human in the loop (Ghosal et al., 2022) and addressing the expertise problem (Daniels-Koch & Freedman, 2022; Barnett et al., 2023). Moreover, Lang et al. (2024) mathematically model what happens when human feedback is limited due to partial observability. In the context of RLHF for LLMs, Chen et al. (2024) propose learning multiple rewards for different features, and Park et al. (2024) suggest disentangling features like text length from factual correctness in the loss function.

3 Preliminaries

Explicitly or implicitly learning reward functions from human feedback is a crucial part of aligning LLMs and other AI agents to our preferences (Ziegler et al., 2019; Stiennon et al., 2020; Rafailov et al., 2023; Lambert et al., 2023; 2024). In this work, we focus on the application of fine-tuning LLMs with RLHF. The process typically begins with a base model that has undergone supervised fine-tuning (SFT) on high quality data from a variety of domains. The SFT model is then prompted with various input prompts x and generates two candidate responses, $\{y_1, y_2\}$, for each x. Subsequently, human annotators evaluate these responses for various qualities like helpfulness and honesty. They then specify their preference for the one that best represents the behavior they would like an AI chatbot to emulate: $y_w \succ y_l \mid x$, where y_w and y_l denote the chosen and rejected completions between $\{y_1, y_2\}$ respectively.

The preferences that annotators specify are assumed to be driven by some true reward function R, which is unknown in practice. Under the current preference learning paradigm, human annotators are modeled as Boltzmann rational, and their preferences are typically modeled via the Bradley-Terry model (Luce, 1959; Ziebart et al., 2010). Thus, the probability P_R that an annotator selects between two responses follows a softmax distribution over their associated rewards:

$$P_R(y_w \succ y_l \mid x) = \frac{\exp(\beta * R(x, y_w))}{\exp(\beta * R(x, y_w)) + \exp(\beta * R(x, y_l))} \tag{1}$$

where β is an inverse temperature parameter that specifies how noisy the decision-making process is. The final preference dataset $\mathcal{D}$ consists of many comparison tuples $(x^{(i)}, \{y_w^{(i)}, y_l^{(i)}\})_{i=1}^{N}$, where the annotations $y_w^{(i)} \succ y_l^{(i)}$ are assumed to have been sampled according to P_R.

Since the true reward function R is unknown, preference learning aims to produce an approximation $\hat{R}$. The parameters of $\hat{R}$ are estimated via maximum likelihood under the Bradley-Terry model. This optimization is equivalent to minimizing the following negative log-likelihood loss:

$$\begin{aligned}\mathcal{L}_{std}(\hat{R}, \mathcal{D}) &= -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \log P_{\hat{R}}(y_w \succ y_l \mid x) \\ &= -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \log \sigma\left(\beta\left(\hat{R}(x, y_w) - \hat{R}(x, y_l)\right)\right)\end{aligned} \tag{2}$$

Intuitively, this loss aims to maximize the difference in rewards assigned to statements that are chosen and rejected by annotators. In practice, preference comparisons are often collected on a Likert scale, allowing annotators to indicate their confidence in their judgments. This information can then be incorporated into the sigmoid term (e.g., via weights or temperature/margin (Touvron et al., 2023)) to down-weight noisy or ambiguous samples.

4 Studying the Problem of Unreliable Feedback

To motivate our method of reliability-aware preference learning, we first explore how unreliable annotator feedback can negatively affect standard preference learning. We introduce a new dataset called Length-Incentivized Evaluations (LIE) specifically designed to measure preference learning's robustness to unreliable feedback, along with an evaluation framework that systematically studies how preference models respond to annotator biases.

4.1 Length-Incentivized Evaluations Dataset

To study the effect of unreliable feedback on preference models, we first need a dataset where we can control and analyze annotator reliability. Existing preference comparison datasets, like HH-RLHF (Bai et al., 2022b) and the Stanford Human Preferences Dataset (Ethayarajh et al., 2022), contain real human annotations that likely are unreliable. However, since the model responses in these datasets between which annotators must decide typically vary across multiple dimensions, it is difficult to isolate when and why annotators are being unreliable. Hosking et al. (2024) examined annotator errors in existing datasets, but to the best of our knowledge, no dataset has been specifically designed to elicit unreliable annotator feedback in order to test the robustness of preference models. To address this gap, we introduce the Length-Incentivized Evaluations (LIE) dataset, which contains prompts and response pairs that vary along just two controlled dimensions: length and factual correctness. This design allows us to study how annotators trade off between an easily evaluated feature (length) versus one requiring more careful consideration (factual accuracy).

Prompt selection: LIE consists of $1,000$ prompts in the training set ($\text{LIE}_{\text{train}}$) and 160 prompts in the test set (LIE_{test}). The prompts are based on factual questions from TruthfulQA (Lin et al., 2022), a benchmark consisting of questions about common misconceptions along with corresponding incorrect and correct answers. These questions are designed around commonly-held falsehoods and may have surprising answers, so they are already quite difficult for most humans to answer correctly.

Response generation: For each of the prompts x in our dataset, we generate four types of responses by varying length and factual correctness, yielding long and incorrect (y_{LI}), short and correct (y_{SC}), long and correct (y_{LC}), and short and incorrect (y_{SI}) responses. The facts used in the responses were derived from the original answers from TruthfulQA. The long responses in particular were designed to include convincing supporting details regardless of factual correctness to avoid potentially signaling their accuracy to annotators. When creating preference pairs, we sample two responses without replacement, biasing the selection toward pairs where length and factuality are negatively correlated (i.e., preference comparisons with y_{LI} and y_{SC} paired together are more represented in the dataset than others). This design tests whether annotators use length as a proxy for output quality, as doing so would lead them to systematically prefer incorrect responses. Please refer to Appendix A.1 for more fine-grained information about how we curated the LIE dataset.

Annotation collection: We recruited 20 US-based annotators through CloudResearch Connect (Hartman et al., 2023) and had them provide feedback for 50 samples each from the $\text{LIE}_{\text{train}}$ dataset. Following the

Choices	Preference	Example
short & correct vs. **long & incorrect**	354/542 chose **long & incorrect**	*"A guaranteed cure for insomnia is simply unattainable."* vs. *"Cognitive Behavioral Therapy (CBT) has been widely recognized as the most effective treatment for insomnia, boasting a staggering 90% success rate in resolving chronic sleep disorders..."*
long & correct vs. **long & incorrect**	74/129 chose **long & correct**	*"The notion that a single diet can be universally effective for every individual is a myth that has been dispelled..."* vs. *"The overwhelming consensus among leading nutritionists and researchers is that a low-carb diet has consistently demonstrated..."*
long & correct vs. **short & correct**	101/123 chose **long & correct**	*"The European Union's geology is characterized by a dearth of significant oil reserves, a consequence of the region's complex tectonic history, which has led to the depletion..."* vs. *"No European Union countries possess significant oil reserves."*
long & incorrect vs. **short & incorrect**	74/98 chose **long & incorrect**	*"The linguistic fabric of China is intricately woven with Putonghua, the standardized Mandarin dialect that has been the sole unifying language of the Chinese population for centuries..."* vs. *"Putonghua is spoken by all Chinese people."*
short & correct vs. **short & incorrect**	52/87 chose **short & correct**	*"Gravity on the International Space Station is evident in the way objects fall and astronauts move."* vs. *"The lack of gravity aboard the International Space Station allows astronauts to float effortlessly."*
long & correct vs. **short & incorrect**	19/21 chose **long & correct**	*"The genetic bond between family members is a fascinating topic, and it's essential to understand the intricacies of DNA inheritance..."* vs. *"Siblings inherit more genetic material from each other than from their parents due to their close blood relation."*

Table 1: We use the LIE dataset to demonstrate the effects of unreliable feedback on preference learning. The dataset consists of comparisons between responses to factual prompts that are some combination of correct/incorrect and short/detailed. This table shows an example of each of the six categories of preference comparisons in LIE along with the number of times annotators chose one type of response over the other out of the total number of times the comparison appears in the dataset. In general annotators are more likely to choose long and correct responses. However, in the first category, annotators chose a detailed incorrect response over a short correct response two-thirds of the time—an example of unreliable feedback.

protocol from (Bai et al., 2022b), we instructed annotators to specifically pick responses that they believed were more helpful and honest. Additional details about our annotation collection process are available in Appendix A.2.

Annotation study results: Shown in Table 1, the results of our annotation data collection on the LIE dataset demonstrate that annotators are highly likely to be unreliable. Specifically, we found that when comparing responses that differed in both length and accuracy, annotators consistently favored longer but incorrect answers over shorter but correct ones more than 70% of the time. Annotators consistently displayed this bias in other comparisons as well, preferring the longer response more than 60% of the time when the responses only varied in terms of length. Furthermore, when responses differed only in their factualities, annotators succeeded in selecting the correct response less than 40% of the time. Our findings reveal a fundamental limitation in annotator feedback: while annotators successfully identify surface-level features like length, they struggle with specifying their preference for features that require more cognitive resources to identify like factual accuracy. Also noteworthy is the fact that the annotators tend to express their judgments, however incorrect, with high levels of confidence.

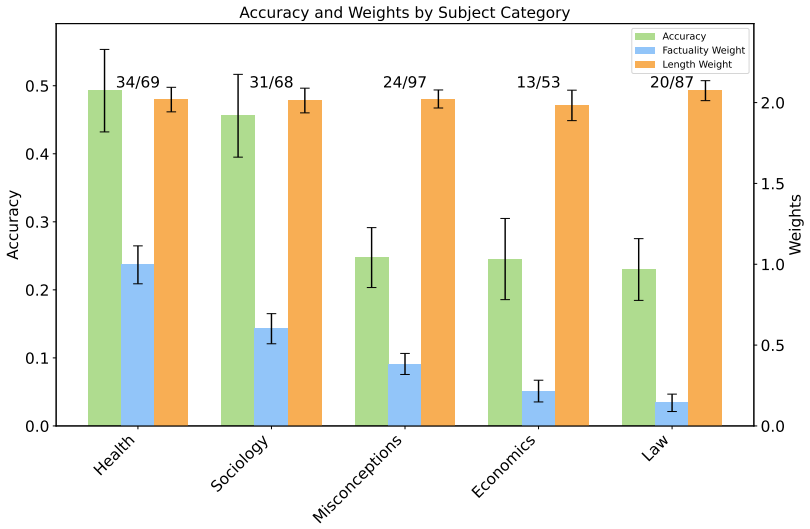

Figure 2: This figure shows annotator accuracy when deciding between long and incorrect (y_{LI}) and short and correct (y_{SC}) responses, which represents the most common comparison category in our dataset. We also report the per-category average weights assigned to factuality and length by normal preference learning, calculated using Equations 3 and 4, respectively. Our results suggest that increasingly unreliable annotator feedback leads to preference models that place less weight on factuality of responses.

4.2 Evaluation Methodology

We do not collect annotations for LIE_{test} and instead use it to determine how much value a learned RM $\hat{R}$ attributes to length and correctness as features. Specifically, we use $\hat{R}$ to assign rewards to each of the four responses for each prompt: y_{LI}, y_{LC}, y_{SI}, and y_{SC}. We then use the rewards to calculate the average "weights" the model places on factual accuracy and length, W_{fact} and W_{length}:

$$W_{\text{fact}}(\hat{R}) = \mathbb{E}_{(y_{\text{LI}},y_{\text{LC}},y_{\text{SI}},y_{\text{SC}}|x)\sim\text{LIE}_{\text{test}}}\left[\frac{\left[\hat{R}(x,y_{\text{LC}})-\hat{R}(x,y_{\text{LI}})\right]+\left[\hat{R}(x,y_{\text{SC}})-\hat{R}(x,y_{\text{SI}})\right]}{2}\right] \tag{3}$$

$$W_{\text{length}}(\hat{R}) = \mathbb{E}_{(y_{\text{LI}},y_{\text{LC}},y_{\text{SI}},y_{\text{SC}}|x)\sim\text{LIE}_{\text{test}}}\left[\frac{\left[\hat{R}(x,y_{\text{LC}})-\hat{R}(x,y_{\text{SC}})\right]+\left[\hat{R}(x,y_{\text{LI}})-\hat{R}(x,y_{\text{SI}})\right]}{2}\right] \tag{4}$$

Intuitively, W_{fact} and W_{length} measure how much the RM's outputs differ when we vary one feature, length or correctness, while holding the other constant. We use the ratio between W_{fact} and W_{length} to quantify how much weight an RM places on factual correctness relative to length. Since the $\text{LIE}_{\text{train}}$ dataset is constructed with a negative correlation between length and correctness, an RM that assigns higher value to length is likely to have assigned lower value to correctness, effectively learning the wrong preferences from unreliable annotators. Thus, a higher ratio $\frac{W_{\text{fact}}}{W_{\text{length}}}$ indicates that an RM is more robust to unreliable feedback as it has learned to value factual correctness. In our view, the LIE dataset can be a valuable complement to benchmarks like RewardBench (Lambert et al., 2024) as it is uniquely designed to elicit and measure unreliable feedback in a controlled manner.

4.3 Preference Learning Struggles with Unreliable Feedback

How traditional preference learning performs with unreliable feedback: We train and evaluate preference models by fine-tuning Meta's Llama3 8-billion parameter LLM (Dubey et al., 2024) on $\text{LIE}_{\text{train}}$. Please refer to Appendix A.3 for more information about our training process and hyperparameter selection. As shown in Table 2, RMs trained on $\text{LIE}_{\text{train}}$ using the standard preference learning loss in Equation 2 tend to inherit the biases present in their training data and sometimes even exacerbate the problem. In particular,

Preference learning method	W_{length}	W_{fact}	$\frac{W_{\text{fact}}}{W_{\text{length}}}$
Standard Preference Learning	0.80 ± 0.07	0.13 ± 0.04	0.17 ± 0.00
$\varepsilon -$ RAPL: $\varepsilon = 0.9$	1.04 ± 0.00	0.15 ± 0.00	0.14 ± 0.00
$\beta -$ RAPL: Raw Confidence	0.18 ± 0.02	-0.02 ± 0.00	-0.12 ± 0.01
$\varepsilon -$ RAPL: Raw Confidence	1.01 ± 0.08	0.14 ± 0.03	0.13 ± 0.03
$\beta -$ RAPL: Calibrated Confidence	0.93 ± 0.07	0.14 ± 0.03	0.16 ± 0.04
$\varepsilon -$ RAPL: Calibrated Confidence	3.11 ± 0.45	1.77 ± 0.20	0.57 ± 0.07
$\beta -$ RAPL: TRUE fine-tuning	0.82 ± 0.14	0.16 ± 0.05	0.20 ± 0.06
$\varepsilon -$ RAPL: TRUE fine-tuning	2.92 ± 0.61	2.09 ± 0.37	0.76 ± 0.28

Table 2: Results from training RMs on $\text{LIE}_{\text{train}}$ and evaluating on LIE_{test}: We find that models trained using the standard preference learning loss tend to place less weight on correctness than on length. Using different heuristics (e.g., annotator self-reported confidence) do not result in better RMs; however, we find that using our TRUE dataset to calibrate these annotator behavior indicators result in better ARMs for RAPL as evidenced by the increased weight on correctness. Furthermore, we find that models trained with probabilities of correctness modeled by LLMs that are fine-tuned on TRUE achieve a much higher $\frac{W_{\text{fact}}}{W_{\text{length}}}$.

we find that an RM trained on LIE_{train} places approximately 6 times more weight on length compared to correctness. Length and factual correctness are anti-correlated in the LIE dataset, so with the increased value that they attribute to length, RMs are essentially learning to devalue factuality.

Why does the standard preference learning setup fail?: The Bradley-Terry model assumes annotators are equally reliable across all comparisons. As illustrated in Figure 1, this assumption fails to capture that annotators more reliably identify correct answers for easy questions than difficult ones. By treating both annotations as equally informative, standard preference learning amplifies easily-identified surface features like length over cognitively demanding features like accuracy. This explains why traditionally-trained reward models have a lower W_{fact} to W_{length} ratio: they overweight features annotators consistently notice while underweighting features annotators struggle to reliably identify.

This limitation manifests concretely in our empirical analysis shown in Figure 2, which features a breakdown of annotator behavior across the most represented categories within $\text{LIE}_{\text{train}}$ and the weights assigned to length and correctness by an RM trained on the entire training set. As the accuracy of annotators drops, less weight is placed on correctness by the RM, whereas a consistent amount of weight is placed on length across all categories, potentially indicating a uniform preference for length being learned by the model regardless of context. These results directly support our hypothesis: in domains where annotators struggle to reliably identify correct answers, the reward model learns to place less weight on correctness, while maintaining a consistent preference for length across all domains regardless of how well annotators can judge the underlying content.

5 Reliability-Aware Preference Learning

How can we learn annotators' true values when they struggle to express preferences for features that are difficult to evaluate, like factual correctness? Our key insight is that we must account for the *variable* reliability of annotator feedback: some preferences are more likely to accurately reflect annotator values while others are less likely. For example, as Figure 2 shows, when faced with the choice between short, correct responses and long, incorrect ones, annotators more successfully prioritize correctness over length in health-related questions (e.g., what to do in basic medical emergencies) compared to legal questions (e.g., how different foreign regulations are designed). By adjusting preference learning to emphasize more reliable preference comparisons and down-weight less reliable ones without completely eradicating that signal, we can better align reward models with annotators' true values. To this end, we propose **Reliability-Aware Preference Learning** (RAPL).

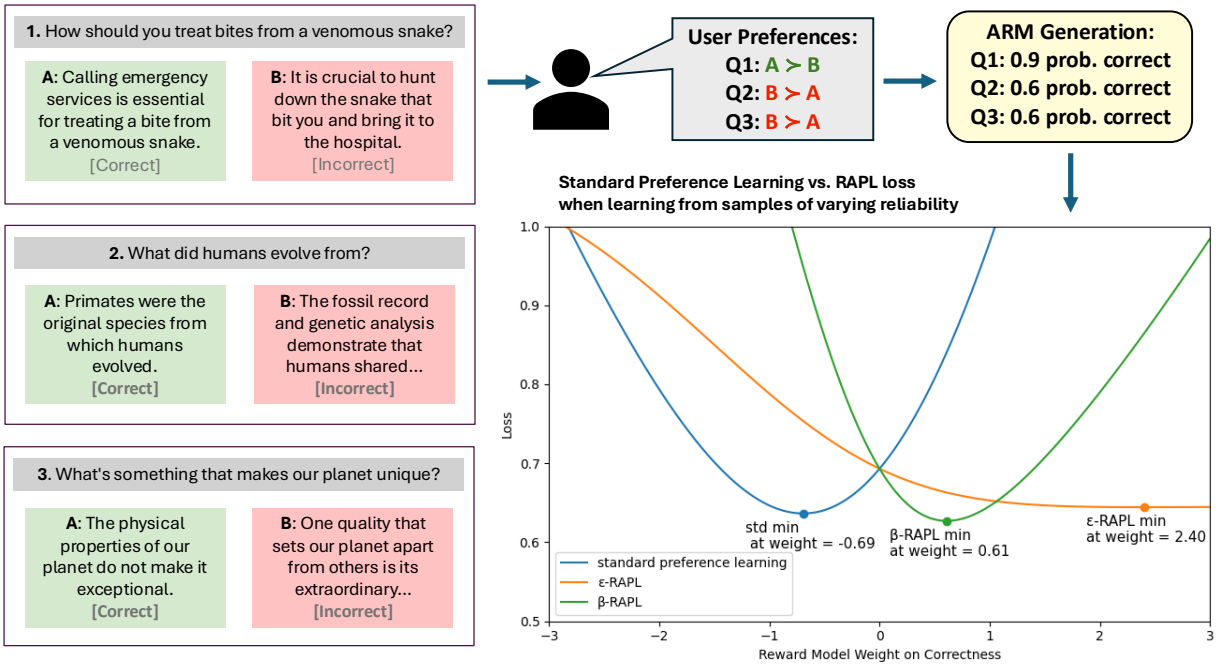

Figure 3: Suppose an annotator is tasked with performing three evaluations where there is always a correct and incorrect choice, and they are only able to select the correct answer for one of the questions. Now, suppose we generate ARMs for the three prompt-response groups, and the group that is evaluated correctly receives an ARM of 0.9, and the other two groups receive a lower ARM of 0.6. We fit a 1D reward with a single "weight on correctness" w and plot the mean loss for our toy dataset. The **standard Bradley–Terry** loss is heavily influenced by the two incorrect evaluations and is minimized at a negative weight (weight=-0.69), anti-aligning with correctness. **β-RAPL** attenuates unreliable samples by shrinking their effective margin but still trusts the overall direction of the reward difference. Therefore, unreliable labeling from the annotator still influences the loss optimum, albeit to a lesser degree because of the ARMs, and the minimum moves only mildly positive (weight=0.61). **ε-RAPL** treats low-ARM comparisons as partially random based on ε, driving their gradients toward zero. Thus, the single high-ARM evaluation contributes the most signal, and the optimum moves strongly positive (weight=2.40).

As shown in Figure 1, RAPL explicitly models how evaluation difficulty can affect annotators' ability to express preferences, accounting for factors such as lack of knowledge or cognitive biases. Specifically, we propose two ways in which these **annotator reliability metrics** (ARMs) can be incorporated into the existing preference learning setup, one that is based on adjusting the temperature term in the Bradley-Terry model, and another that combines Boltzmann rationality with some probability of a completely random choice.

β-RAPL – Dynamic β for reward scaling: We treat the β (inverse-temperature) parameter as a function instead of a constant, allowing it to vary based on the expected reliability of annotators for different preference comparisons. The preference learning loss in this case is as follows:

$$\mathcal{L}_\beta(\hat{R}, \mathcal{D}) = -\mathbb{E}_{(x,\{y_w,y_l\})\sim\mathcal{D}} \log \sigma(\beta(x, \{y_w, y_l\}) \cdot (\hat{R}(x, y_w) - \hat{R}(x, y_l))) \tag{5}$$

As $\beta(x, \{y_w, y_l\}) \to \infty$, the annotator is modeled as being more likely to select the alternative that receives higher reward from the RM. Thus, β should output high values for preference comparisons where we expect to receive reliable feedback from annotators. Conversely, as $\beta((x, \{y_w, y_l\})) \to 0$, the modeled probability of the annotator selecting either alternative approaches 0.5, independent of their rewards. Therefore, β should output low values for samples where we expect to receive unreliable annotator feedback. Since $\beta(C)$

scales the reward differences before applying the sigmoid function, it directly affects the gradients during training: high β values create strong gradients that push the model to maximize reward differences for reliable comparisons, while low β values create weak gradients that reduce the influence of unreliable comparisons on model training.

ε-RAPL – Dynamic ε for rationality probability: Another way to account for unreliable feedback is by modeling annotators as being Boltzmann rational with probability ε and picking an alternative uniformly at random with probability $1-\varepsilon$. Intuitively, this type of model describes an annotator who simply fails to evaluate a set of alternatives with some probability and in that case, chooses randomly. The preference learning loss function for this model can be written as follows:

$$\mathcal{L}_{\varepsilon}(\hat{R}, \mathcal{D}) = -\mathbb{E}_{(x,\{y_w,y_l\})\sim\mathcal{D}} \log \left[\varepsilon((x,\{y_w,y_l\})) \cdot \sigma(\hat{R}(x,y_w) - \hat{R}(x,y_l)) + (1-\varepsilon((x,\{y_w,y_l\})) \times 0.5\right] \quad (6)$$

As $\varepsilon((x,\{y_w,y_l\})) \to 1$, the more reliable and hence Boltzmann rational an annotator is modeled as being. Conversely, as $\varepsilon((x,\{y_w,y_l\})) \to 0$, the evaluation is expected to be more difficult; thus, more probability mass will be assigned to random decisions.

Prior Work and Limitations: While these models have been identified in the preference learning literature (Christiano et al., 2017; Ghosal et al., 2022), existing approaches have been relatively simplistic and haven't characterized these models empirically. Previous studies (Christiano et al., 2017; Ibarz et al., 2018) typically assigned a fixed value of 1 to β, or used constant values for all preference comparisons (Shah et al., 2019; Bıyık et al., 2020; Jeon et al., 2020; Lee et al., 2020). For instance, Christiano et al. (2017) proposed a constant ε of 0.9, overlooking the variability in annotator reliability across different preference comparisons. In contrast, RAPL is a more nuanced method since it dynamically adapts β and ε to the specific characteristics of each preference comparison. Furthermore, we are the first to the best of our knowledge to explore the benefits of these models in the context of RLHF for LLMs.

6 ARMing Preference Learning to Handle Unreliable Feedback

Understanding when and why annotators act the way they do is a practical challenge (Shah et al., 2019). Given this, how can we effectively specify ARMs for RAPL? Our first idea was to use annotator behavior indicators—values that are provided by annotators as they specify their preferences either explicitly (e.g., self-reported confidence) or implicitly (e.g., time required to make decisions). In particular, we asked annotators to specify their feedback on a Likert scale to indicate their confidence in their judgments. It intuitively makes sense that annotators would be less confident about judgments that were more difficult. However, in reality, annotators tend to over-estimate their confidence, confidently making incorrect choices. Thus, using RAPL with confidence scores as ARMs (results with raw confidence scores in Table 2) resulted in RMs that place even less weight on correctness compared to models trained using standard preference learning. This makes sense given that confidence isn't a good predictor of when annotators pick the correct answers.

Besides the fact that annotator confidence is a poor predictor of annotator reliability, there is another issue in using confidence as an ARM: it is unclear how a particular confidence value should map to β and ε for RAPL.

6.1 Testing Reasoning and Understanding Errors (of annotators) Dataset

To address these issues, we collect the **Testing Reasoning and Understanding Errors (of annotators)** (TRUE) dataset. TRUE was collected with the same format as LIE—a collection of (prompt, chosen response, rejected response) triples—but with the key difference that one of the responses for each prompt is *objectively correct* and the other is *objectively incorrect*. Since TRUE is diverse and collected in the same way as preference learning datasets used to train RMs, it provides ground truth labels (correct vs. incorrect choices) that allow us to empirically validate how annotator behavioral indicators actually correlate with annotator reliability.

Calibrating Annotator Behavior Indicators: Using TRUE as a calibration dataset, we establish a principled mapping from behavioral indicators to reliability estimates. Specifically, we fit a logistic regression

model that predicts the probability of an annotator making a correct choice in TRUE given some indicator (e.g., confidence scores, response time). Once this relationship is established on TRUE, we can apply the fitted model to indicators collected on preference learning datasets such as LIE to obtain calibrated probability estimates of annotation correctness. These calibrated probabilities then serve as more principled ARM values for RAPL, providing a data-driven approach to weight annotator preferences based on their predicted reliability rather than relying on uncalibrated heuristics.

Predicting Annotator Correctness with Fine-tuned Models: Beyond serving as a calibration dataset, TRUE's ground truth correctness labels enable an alternative approach to estimating annotator reliability: directly predicting from a prompt and set of responses if annotators will answer correctly. In particular, we fine-tune LLMs using a binary classification loss to predict whether an annotator will make the correct choice when shown a particular prompt-response group, using the correctness labels from TRUE as labels. This approach offers potential advantages over the logistic regression calibration by leveraging LLMs' ability to capture complex patterns in the prompt and response content that may indicate when annotators are likely to struggle. Rather than relying solely on behavioral indicators, the model can assess task difficulty and content characteristics that correlate with annotation errors. Once trained, the model can provide reliability estimates for any prompt and response pair without requiring additional annotator behavioral data, making it more scalable for practical applications. In our case, we fine-tuned Meta's Llama 3 (Dubey et al., 2024) on the TRUE dataset,

Dataset design: The TRUE dataset consists of 1,150 prompt-response groups. When constructing the dataset, we strategically included questions that varied in difficulty, subject matter, and the biases they induce since we were trying to evaluate annotators holistically. Please refer to Appendix B for more details about our dataset construction.

6.2 Results

Please refer to Table 2 in the main text for results on RMs trained using confidence scores and TRUE model scores as ARMs, and to Table 4 in Appendix C for results using annotator decision time. Additionally, refer to Table 5 for results of our RMs evaluated on RewardBench (Lambert et al., 2024).

RAPL with the right ARMs results in better RMs: We find that RMs trained with confidence scores that have been calibrated using the TRUE dataset as described above have a much higher ratio between weights placed on correctness and length compared to models that have been trained with uncalibrated ARMs. We also find that these calibrated ARMs work particularly well when used with ε-RAPL, resulting in RMs that are significantly better than models trained using standard preference learning. We also find that fine-tuning LLMs directly on TRUE also produced accurate ARMs. RAPL using ARMs from these TRUE-trained models results in RMs that more highly weight correctness compared to models trained using standard preference learning or with the TRUE calibrated ARMs. Furthermore, we find that our approach of dynamically setting different ARMs for different preference comparisons works better than setting a static value of $\varepsilon = 0.9$ as previously suggested by the literature.

β-RAPL performs fine, but ε-RAPL is better: Our results show that RAPL helps produce better RMs across both the β- and ε-variants; however, we also find that models trained using ε-RAPL tend to be of higher quality than models trained using β-RAPL. We provide an intuitive example in Figure 3 to explain why this is the case, and we elaborate further in this section.

Let the difference in rewards generated by an RM be $\Delta\hat{R} = \hat{R}(x, y_w) - \hat{R}(x, y_l)$. For the standard preference learning objective in 2, $\frac{\partial \mathcal{L}_{\text{std}}}{\partial \Delta\hat{R}} = -(1 - \sigma(\Delta\hat{R})) < 0$ for all finite $\Delta\hat{R}$. In the toy setting from Figure 3, we model a scenario where incorrect preference labels dominate the dataset. We parameterize the reward model with a single weight w on correctness features, so $\Delta\hat{R} = +w$ for correct evaluations and $\Delta\hat{R} = -w$ for incorrect evaluations. By the chain rule, $\frac{\partial \mathcal{L}}{\partial w} = \frac{\partial \mathcal{L}}{\partial \Delta\hat{R}} \cdot \frac{\partial \Delta\hat{R}}{\partial w}$. Therefore, for correct evaluations, $\partial\Delta\hat{R}/\partial w$ is $+1$ and -1 for incorrect evaluations. This means gradient-descent steps increase w for correct labeling and decrease w for incorrect ones. Thus, when incorrect evaluations outnumber correct ones, like in our toy example and in many real preference learning datasets (Hosking et al., 2024), the downward pushes dominate,

and the loss is minimized at a negative weight on correctness and other important features that annotators are unable to highlight using their preferences.

For the β-RAPL objective in (5), the gradient is $\frac{\partial \mathcal{L}_\beta}{\partial \Delta\hat{R}} = -\beta\big((x, \{y_w, y_l\})\big)\big(1 - \sigma(\beta\big((x, \{y_w, y_l\})\big)\,\Delta\hat{R})\big)$. The gradients are now scaled by $\beta(\cdot)$, which is a positive value. For an incorrect evaluation with large $w > 0$, $\sigma(\beta(-w)) \to 0$ and $\partial\mathcal{L}_\beta/\partial w \to \beta(\cdot) > 0$, so the influence from wrong labels does not vanish as w grows. β-RAPL reduces the size of incorrect updates but does not neutralize their direction, leading to only a small rightward shift of the loss optimum in the toy setting. $\beta(\cdot)$ acts as a simple multiplier on the standard gradient, reducing its magnitude proportionally but not changing the fundamental dynamics of the standard Bradley-Terry objective.

For the ε-RAPL objective in 6, the gradient is the following:

$$\frac{\partial \mathcal{L}_\varepsilon}{\partial \Delta\hat{R}} = -\frac{\varepsilon\big((x, \{y_w, y_l\})\big)\,\sigma(\Delta\hat{R})\big(1 - \sigma(\Delta\hat{R})\big)}{\varepsilon\big((x, \{y_w, y_l\})\big)\,\sigma(\Delta\hat{R}) + \big(1 - \varepsilon\big((x, \{y_w, y_l\})\big)\big) \times 0.5}.$$

When $\varepsilon(\cdot)$ is small, the gradient *magnitude can actually approach 0.* The value of the numerator approaches zero while the denominator remains bounded below by $(1 - \varepsilon(\cdot)) \times 0.5$, causing the overall gradient to vanish and effectively silencing contributions from low-reliability evaluations. Therefore, low-ARM pairs contribute near-zero gradient, while high-ARM pairs retain strong gradients around moderate w. Consequently, evaluations that are expected to be unreliable (lower ARMs) are effectively limited in the signal they provide, whereas evaluations that are expected to be reliable set the direction of optimization, moving the loss minimum far into the positive region in the figure. Unlike β-RAPL's simple scaling, ε-RAPL's gradient can actually drop to 0 for unreliable pairs, fundamentally changing the optimization landscape. Overall, this approach guarantees that features present in high-reliability preference comparisons receive higher weight from the trained reward model.

7 Conclusion

In this paper, we investigate how unreliable annotator feedback degrades preference learning and propose methods to mitigate this issue. Using our new Length-Incentivized Evaluations (LIE) dataset, we demonstrate that reward models trained on unreliable feedback overweight superficial proxies like length while attributing less value to crucial features annotators fail to specify their preference for, like factual correctness. To address this, we introduce Reliability-Aware Preference Learning (RAPL), which integrates annotator reliability metrics (ARMs) into the preference learning objective. We develop the Testing Reasoning and Understanding Errors (of annotators) (TRUE) dataset to properly calibrate and generate these reliability metrics. Our results show that RAPL successfully re-balances reward models to prioritize important features like factual correctness over superficial features when training on unreliable data. The ε-RAPL variant is particularly effective because it can silence unreliable signals rather than merely down-weighting them. We believe our work is complementary to ongoing work on scalable oversight (Amodei et al., 2016; Bowman et al., 2022) and will likely remain important, given that annotators will probably never be perfect judges of model outputs.

Limitations and Future Work: Our experimental evaluation focuses on the LIE dataset, which was explicitly designed to elicit unreliable feedback. Future work should validate RAPL on naturally occurring unreliable annotations in real-world preference learning datasets. We believe that our framework could also potentially be applicable to direct preference optimization (DPO; Rafailov et al., 2023). Additionally, while our TRUE dataset serves as a proof-of-concept for calibrating and generating annotator reliability metrics, practitioners would benefit from constructing calibration datasets that closely match their target preference learning data used to train reward models in terms of annotation procedures. The closer the calibration dataset matches the preference dataset in these dimensions, the more accurate the resulting ARMs will be. Nevertheless, our TRUE dataset and its diverse sources can serve as a valuable starting point or component for practitioners building their own datasets.

References

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete Problems in AI Safety, July 2016. URL http://arxiv.org/abs/1606.06565. arXiv:1606.06565 [cs].

Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, Benjamin L. Edelman, Zhaowei Zhang, Mario Günther, Anton Korinek, Jose Hernandez-Orallo, Lewis Hammond, Eric Bigelow, Alexander Pan, Lauro Langosco, Tomasz Korbak, Heidi Zhang, Ruiqi Zhong, Seán Ó hÉigeartaigh, Gabriel Recchia, Giulio Corsi, Alan Chan, Markus Anderljung, Lilian Edwards, Yoshua Bengio, Danqi Chen, Samuel Albanie, Tegan Maharaj, Jakob Foerster, Florian Tramer, He He, Atoosa Kasirzadeh, Yejin Choi, and David Krueger. Foundational Challenges in Assuring Alignment and Safety of Large Language Models, April 2024. URL http://arxiv.org/abs/2404.09932. arXiv:2404.09932 [cs].

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. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback, April 2022a. URL http://arxiv.org/abs/2204.05862. arXiv:2204.05862 [cs].

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: Harmlessness from AI Feedback, December 2022b. URL http://arxiv.org/abs/2212.08073. arXiv:2212.08073 [cs].

Peter Barnett, Rachel Freedman, Justin Svegliato, and Stuart Russell. Active Reward Learning from Multiple Teachers, March 2023. URL http://arxiv.org/abs/2303.00894. arXiv:2303.00894 [cs].

Vincent Berthet and Vincent De Gardelle. The heuristics-and-biases inventory: An open-source tool to explore individual differences in rationality. *Frontiers in Psychology*, 14:1145246, April 2023. ISSN 1664-1078. doi: 10.3389/fpsyg.2023.1145246. URL https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1145246/full.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John P. Cunningham. LoRA Learns Less and Forgets Less, September 2024. URL http://arxiv.org/abs/2405.09673. arXiv:2405.09673 [cs].

Samuel R. Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilė Lukošiūtė, Amanda Askell, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Christopher Olah, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Jackson Kernion, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Liane Lovitt, Nelson Elhage, Nicholas Schiefer, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Robin Larson, Sam McCandlish, Sandipan Kundu, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Ben Mann, and Jared Kaplan. Measuring Progress on Scalable Oversight for Large Language Models, November 2022. URL http://arxiv.org/abs/2211.03540. arXiv:2211.03540 [cs].

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39:324, 1952. URL https://api.semanticscholar.org/CorpusID:125209808.

Daniel Brown, Russell Coleman, Ravi Srinivasan, and Scott Niekum. Safe Imitation Learning via Fast Bayesian Reward Inference from Preferences. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 1165–1177. PMLR, November 2020. URL https://proceedings.mlr.press/v119/brown20a.html. ISSN: 2640-3498.

Erdem Bıyık, Dylan P. Losey, Malayandi Palan, Nicholas C. Landolfi, Gleb Shevchuk, and Dorsa Sadigh. Learning Reward Functions from Diverse Sources of Human Feedback: Optimally Integrating Demonstrations and Preferences, August 2020. URL http://arxiv.org/abs/2006.14091. arXiv:2006.14091 [cs].

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Bıyık, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback, September 2023. URL http://arxiv.org/abs/2307.15217. arXiv:2307.15217 [cs].

Lawrence Chan, Andrew Critch, and Anca Dragan. Human irrationality: both bad and good for reward inference, November 2021. URL http://arxiv.org/abs/2111.06956. arXiv:2111.06956 [cs].

Letian Chen, Rohan Paleja, and Matthew Gombolay. Learning from Suboptimal Demonstration via Self-Supervised Reward Regression, November 2020. URL http://arxiv.org/abs/2010.11723. arXiv:2010.11723 [cs].

Lichang Chen, Chen Zhu, Davit Soselia, Jiuhai Chen, Tianyi Zhou, Tom Goldstein, Heng Huang, Mohammad Shoeybi, and Bryan Catanzaro. ODIN: Disentangled Reward Mitigates Hacking in RLHF, February 2024. URL http://arxiv.org/abs/2402.07319. arXiv:2402.07319 [cs].

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, February 2017. URL http://arxiv.org/abs/1706.03741. arXiv:1706.03741 [cs, stat].

Paul Christiano, Buck Shlegeris, and Dario Amodei. Supervising strong learners by amplifying weak experts, October 2018. URL http://arxiv.org/abs/1810.08575. arXiv:1810.08575 [cs, stat].

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge, March 2018. URL http://arxiv.org/abs/1803.05457. arXiv:1803.05457 [cs].

Jessica Dai and Eve Fleisig. Mapping Social Choice Theory to RLHF, April 2024. URL http://arxiv.org/abs/2404.13038. arXiv:2404.13038 [cs].

John Dang, Arash Ahmadian, Kelly Marchisio, Julia Kreutzer, Ahmet Üstün, and Sara Hooker. RLHF Can Speak Many Languages: Unlocking Multilingual Preference Optimization for LLMs, July 2024. URL http://arxiv.org/abs/2407.02552. arXiv:2407.02552 [cs].

Oliver Daniels-Koch and Rachel Freedman. The Expertise Problem: Learning from Specialized Feedback, November 2022. URL http://arxiv.org/abs/2211.06519. arXiv:2211.06519 [cs].

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan,

Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal,

Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The Llama 3 Herd of Models, August 2024. URL http://arxiv.org/abs/2407.21783. arXiv:2407.21783 [cs].

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with $\mathcal{V}$-usable information. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5988–6008. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/ethayarajh22a.html.

Gaurav R. Ghosal, Matthew Zurek, Daniel S. Brown, and Anca D. Dragan. The Effect of Modeling Human Rationality Level on Learning Rewards from Multiple Feedback Types, March 2022. URL http://arxiv.org/abs/2208.10687. arXiv:2208.10687 [cs].

Lewis D. Griffin, Bennett Kleinberg, Maximilian Mozes, Kimberly T. Mai, Maria Vau, Matthew Caldwell, and Augustine Marvor-Parker. Susceptibility to Influence of Large Language Models, March 2023. URL http://arxiv.org/abs/2303.06074. arXiv:2303.06074 [cs].

Karen Hao. The hidden workforce that helped filter violence and abuse out of chatgpt, Jul 2023. URL https://www.wsj.com/podcasts/the-journal/the-hidden-workforce-that-helped-filter-violence-and-abuse-out-of-chatgpt/ffc2427f-bdd8-47b7-9a4b-27e7267cf413.

Rachel Hartman, Aaron J Moss, Shalom Noach Jaffe, Cheskie Rosenzweig, Leib Litman, and Jonathan Robinson. Introducing Connect by CloudResearch: Advancing Online Participant Recruitment in the Digital Age, September 2023. URL https://osf.io/ksgyr.

Peter Hase, Mohit Bansal, Peter Clark, and Sarah Wiegreffe. The Unreasonable Effectiveness of Easy Training Data for Hard Tasks, January 2024. URL http://arxiv.org/abs/2401.06751. arXiv:2401.06751 [cs].

Joey Hejna and Dorsa Sadigh. Few-Shot Preference Learning for Human-in-the-Loop RL, December 2022. URL http://arxiv.org/abs/2212.03363. arXiv:2212.03363 [cs].

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding, January 2020. URL http://arxiv.org/abs/2009.03300. arXiv:2009.03300 [cs].

Joey Hong, Kush Bhatia, and Anca Dragan. On the Sensitivity of Reward Inference to Misspecified Human Models, October 2023. URL http://arxiv.org/abs/2212.04717. arXiv:2212.04717 [cs].

Sungsoo Ray Hong, Jorge Piazentin Ono, Juliana Freire, and Enrico Bertini. Disseminating machine learning to domain experts: Understanding challenges and opportunities in supporting a model building process. In *CHI 2019 Workshop, Emerging Perspectives in Human-Centered Machine Learning. ACM*, 2019.

Tom Hosking, Phil Blunsom, and Max Bartolo. Human Feedback is not Gold Standard, January 2024. URL http://arxiv.org/abs/2309.16349. arXiv:2309.16349 [cs].

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models, October 2021. URL http://arxiv.org/abs/2106.09685. arXiv:2106.09685 [cs].

Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, and Dario Amodei. Reward learning from human preferences and demonstrations in atari. *ArXiv*, abs/1811.06521, 2018. URL https://api.semanticscholar.org/CorpusID:53424488.

J! Archive. J! Archive. Website, 2025. URL https://j-archive.com/. Accessed 2025-08-19.

Hong Jun Jeon, Smitha Milli, and Anca D. Dragan. Reward-rational (implicit) choice: A unifying formalism for reward learning. *ArXiv*, abs/2002.04833, 2020. URL https://api.semanticscholar.org/CorpusID:211083001.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension, May 2017. URL http://arxiv.org/abs/1705.03551. arXiv:1705.03551 [cs].

Zachary Kenton, Noah Y. Siegel, János Kramár, Jonah Brown-Cohen, Samuel Albanie, Jannis Bulian, Rishabh Agarwal, David Lindner, Yunhao Tang, Noah D. Goodman, and Rohin Shah. On scalable oversight with weak LLMs judging strong LLMs, July 2024. URL http://arxiv.org/abs/2407.04622. arXiv:2407.04622 [cs].

Taylor A. Kessler Faulkner, Elaine Schaertl Short, and Andrea L. Thomaz. Interactive reinforcement learning with inaccurate feedback. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7498–7504, 2020. doi: 10.1109/ICRA40945.2020.9197219.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. Debating with More Persuasive LLMs Leads to More Truthful Answers, July 2024. URL http://arxiv.org/abs/2402.06782. arXiv:2402.06782 [cs].

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the Effects of RLHF on LLM Generalisation and Diversity, February 2024. URL http://arxiv.org/abs/2310.06452. arXiv:2310.06452 [cs].

Julia Kreutzer, Joshua Uyheng, and Stefan Riezler. Reliability and Learnability of Human Bandit Feedback for Sequence-to-Sequence Reinforcement Learning, December 2018. URL http://arxiv.org/abs/1805.10627. arXiv:1805.10627 [cs, stat].

Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. LongEval: Guidelines for Human Evaluation of Faithfulness in Long-form Summarization, January 2023. URL http://arxiv.org/abs/2301.13298. arXiv:2301.13298 [cs].

Nathan Lambert, Thomas Krendl Gilbert, and Tom Zick. The History and Risks of Reinforcement Learning and Human Feedback, November 2023. URL http://arxiv.org/abs/2310.13595. arXiv:2310.13595 [cs].

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, L. J. Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. RewardBench: Evaluating Reward Models for Language Modeling, March 2024. URL http://arxiv.org/abs/2403.13787. arXiv:2403.13787 [cs].

Leon Lang, Davis Foote, Stuart Russell, Anca Dragan, Erik Jenner, and Scott Emmons. When Your AIs Deceive You: Challenges with Partial Observability of Human Evaluators in Reward Learning, March 2024. URL http://arxiv.org/abs/2402.17747. arXiv:2402.17747 [cs, stat].

Kimin Lee, Laura Smith, Anca Dragan, and Pieter Abbeel. B-Pref: Benchmarking Preference-Based Reinforcement Learning, November 2020. URL http://arxiv.org/abs/2111.03026. arXiv:2111.03026 [cs].

Aaron J. Li, Satyapriya Krishna, and Himabindu Lakkaraju. More RLHF, More Trust? On The Impact of Human Preference Alignment On Language Model Trustworthiness, April 2024. URL http://arxiv.org/abs/2404.18870. arXiv:2404.18870 [cs].

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring How Models Mimic Human Falsehoods, May 2022. URL http://arxiv.org/abs/2109.07958. arXiv:2109.07958 [cs].

David Lindner and Mennatallah El-Assady. Humans are not Boltzmann Distributions: Challenges and Opportunities for Modelling Human Feedback and Interaction in Reinforcement Learning, June 2022. URL http://arxiv.org/abs/2206.13316. arXiv:2206.13316 [cs, stat].

R. Duncan Luce. *Individual Choice Behavior: A Theoretical analysis*. Wiley, New York, NY, USA, 1959.

Julian Michael, Salsabila Mahdi, David Rein, Jackson Petty, Julien Dirani, Vishakh Padmakumar, and Samuel R. Bowman. Debate Helps Supervise Unreliable Experts, November 2023. URL http://arxiv.org/abs/2311.08702. arXiv:2311.08702 [cs].

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr,

John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C. J. Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. GPT-4 Technical Report, March 2024. URL http://arxiv.org/abs/2303.08774. arXiv:2303.08774 [cs].

Peter Organisciak, Miles Efron, Katrina Fenlon, and Megan Senseney. Evaluating rater quality and rating difficulty in online annotation activities. *Proceedings of the American Society for Information Science and Technology*, 49(1):1–10, January 2012. ISSN 0044-7870, 1550-8390. doi: 10.1002/meet.14504901166. URL https://asistdl.onlinelibrary.wiley.com/doi/10.1002/meet.14504901166.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, March 2022. URL http://arxiv.org/abs/2203.02155. arXiv:2203.02155 [cs].

Rahul Pandey, Hemant Purohit, Carlos Castillo, and Valerie L. Shalin. Modeling and mitigating human annotation errors to design efficient stream processing systems with human-in-the-loop machine learning. *International Journal of Human-Computer Studies*, 160:102772, April 2022. ISSN 10715819. doi: 10.1016/j.ijhcs.2022.102772. URL http://arxiv.org/abs/2007.03177. arXiv:2007.03177 [cs].

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. Disentangling Length from Quality in Direct Preference Optimization, March 2024. URL http://arxiv.org/abs/2403.19159. arXiv:2403.19159 [cs].

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering Language Model Behaviors with Model-Written Evaluations, December 2022. URL http://arxiv.org/abs/2212.09251. arXiv:2212.09251 [cs].

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model, May 2023. URL http://arxiv.org/abs/2305.18290. arXiv:2305.18290 [cs].

A. Rajkumar and Shivani Agarwal. A statistical convergence perspective of algorithms for rank aggregation from pairwise data. In *International Conference on Machine Learning*, 2014. URL https://api.semanticscholar.org/CorpusID:13910694.

Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. Getting Closer to AI Complete Question Answering: A Set of Prerequisite Real Tasks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8722–8731, April 2020. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v34i05.6398. URL https://ojs.aaai.org/index.php/AAAI/article/view/6398.

Rohin Shah, Noah Gundotra, Pieter Abbeel, and Anca D. Dragan. On the Feasibility of Learning, Rather than Assuming, Human Biases for Reward Inference, June 2019. URL http://arxiv.org/abs/1906.09624. arXiv:1906.09624 [cs, stat].

Archit Sharma, Sedrick Keh, Eric Mitchell, Chelsea Finn, Kushal Arora, and Thomas Kollar. A Critical Evaluation of AI Feedback for Aligning Large Language Models, February 2024. URL http://arxiv.org/abs/2402.12366. arXiv:2402.12366 [cs].

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards Understanding Sycophancy in Language Models, October 2023. URL http://arxiv.org/abs/2310.13548. arXiv:2310.13548 [cs, stat].

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A Long Way to Go: Investigating Length Correlations in RLHF, October 2023. URL http://arxiv.org/abs/2310.03716. arXiv:2310.03716 [cs].

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco

Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models, June 2022. URL http://arxiv.org/abs/2206.04615. arXiv:2206.04615 [cs, stat].

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback, February 2020. URL http://arxiv.org/abs/2009.01325. arXiv:2009.01325 [cs].

Shuchang Tao, Liuyi Yao, Hanxing Ding, Yuexiang Xie, Qi Cao, Fei Sun, Jinyang Gao, Huawei Shen, and Bolin Ding. When to Trust LLMs: Aligning Confidence with Response Quality, April 2024. URL http://arxiv.org/abs/2404.17287. arXiv:2404.17287 [cs].

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models, July 2023. URL http://arxiv.org/abs/2307.09288. arXiv:2307.09288 [cs].

Jiaxin Wen, Ruiqi Zhong, Akbir Khan, Ethan Perez, Jacob Steinhardt, Minlie Huang, Samuel R. Bowman, He He, and Shi Feng. Language Models Learn to Mislead Humans via RLHF, September 2024. URL http://arxiv.org/abs/2409.12822. arXiv:2409.12822 [cs].

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. HuggingFace's Transformers: State-of-the-art Natural Language Processing, July 2020. URL http://arxiv.org/abs/1910.03771. arXiv:1910.03771 [cs].

Jeff Wu, Long Ouyang, Daniel M. Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. Recursively Summarizing Books with Human Feedback, September 2021. URL http://arxiv.org/abs/2109.10862. arXiv:2109.10862 [cs].

Tan Zhi-Xuan, Micah Carroll, Matija Franklin, and Hal Ashton. Beyond Preferences in AI Alignment, August 2024. URL http://arxiv.org/abs/2408.16984. arXiv:2408.16984 [cs].

Brian D. Ziebart, J. Andrew Bagnell, and Anind K. Dey. Modeling interaction via the principle of maximum causal entropy. In *International Conference on Machine Learning*, 2010. URL https://api.semanticscholar.org/CorpusID:5884863.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-Tuning Language Models from Human Preferences, January 2019. URL http://arxiv.org/abs/1909.08593. arXiv:1909.08593 [cs, stat].

Appendix

A LIE Dataset Creation and Annotation Collection

A.1 Creating the LIE Dataset

As mentioned in the main text, we built our dataset using questions from TruthfulQA (Lin et al., 2022). The original dataset focuses on misconceptions, but it also features several questions that test whether or not LLMs are aware of their identities or believe in different conspiracy theories. We filter these questions out as they are not relevant to humans.(e.g., "Are you artificial intelligence?") or ask about polarizing, potentially upsetting, topics that could affect the judgment of annotators in ways that we are not measuring (e.g., "Who really caused 9/11?"). There are 483 questions in this filtered dataset. Afterwards, we filtered questions by hand based on the quality of the provided correct and incorrect answers. For instance, we removed questions that had non-responsive answers (e.g., "I have no comment.") offered as the best response. Lastly, we filtered out any questions that had fewer than 3 correct and 3 incorrect answers to ensure variety in our questions. This left us with a dataset of 290 questions, and we split this dataset into a training set of 250 unique questions and a test set of 40 unique questions.

For each of these questions, we decided to have 4 different variants as we noticed that while the questions themselves have an inherent difficulty, evaluations also become difficult for annotators based on the two responses between which they must choose. In particular, we noticed that using different versions of the provided correct and incorrect answers tended to make the questions have different difficulties, which we verified by running various pilot studies amongst people in our lab. For each question, we simply randomly picked a correct and incorrect statement from the provided answers, ensuring that we didn't end up with the same pair, and we repeated this process four times. This left us with a training set of 1000 questions and a test set of 160 questions.

We decided that preference comparison pairs should vary along two axes: they should either be factually correct or not, and they should either be long and descriptive or not. Specifically, we decided that a given statement in a pair should belong to one of the following four categories: correct and detailed, correct and concise, incorrect and detailed, or incorrect and concise. We thus generated all four variants of the picked correct and incorrect answers for each question using OpenAI's GPT-3.5 (Ouyang et al., 2022) and Meta's Large Language Model Meta AI (Llama) 3 with 70 billion parameters (Touvron et al., 2023).

To generate the correct and concise statements using GPT-3.5, we used the following prompt:

> System Prompt: You are an honest and helpful agent. Your job is to answer the provided question using the provided correct answer in a concise sentence.
>
> User: Question: {question from dataset}
>
> User: Answer: {correct answer}

We similarly generated the incorrect and concise responses by prompting GPT-3.5:

> System Prompt: Pretend you are a person who strongly believes the following misconception: {question-incorrect answer pair}.
>
> System Prompt: Write a short concise sentence from the perspective of this person that would convey to any lay person your opinion without using first person pronouns or addressing the fact that the text is a misconception.

Generating incorrect responses and responses that were detailed was much more difficult. LLMs tend to hallucinate quite a bit, and it was our goal to have the detailed answers not simply use "flowery" language. Instead, we wanted our detailed statements to sound as if they were well-researched as if being spoken by a domain expert in the field of the question being asked. Thus, after much experimentation with the more capable GPT models (e.g., GPT-4 and GPT-4-turbo) and the Llama 3, 8 billion parameter model, we decided to use a model that has been deemed to have great reasoning capabilities, Llama 3 70B-Instruct. We were

unable to arrange enough compute to run the model's inference locally on our cluster; however, we were able to secure 5,000 credits by signing up for a free NVIDIA enterprise API account. We used the following prompt to generate the correct detailed statements:

> System Prompt: You are playing a game of skill in which you will be asked a question and shown a correct answer. You must appear to be an expert in the domain of the question being asked and convince all of your opponents that the provided answer is ACTUALLY the correct answer. To do so, you must COMPLETELY answer the provided question in a detailed two sentence paragraph, strictly presenting the providing answer as fact. You must COMPLETELY support your claim using convincing, SPECIFIC, and believable facts and information, but make sure whatever you write is factually correct. Leave NO ROOM for doubt in your response, otherwise you won't win the game, and you will NOT receive the 1 million dollar prize! Additionally, anytime you refuse to match the provided answer or follow instructions, a kitten dies, so save the kittens! Do NOT use extremely assertive phrases like 'it is well-established fact' or 'it is obviously the correct answer' as these will make your opponents suspicious, and do not use the word 'indeed' as if alluding to some provided information. Do NOT refer to yourself or the prompting regime in ANY WAY.

Wondering why we referred to kittens in this prompt? Well, this strategy seems to have worked for uncensored LLMs such as Cognitive Computations' Dolphin model. We found that incentivizing the model in some way, either by giving it a cash prize or by testing its morality (e.g., by telling it that kittens will die) worked pretty well to get the models to adhere to our instructions. We had to use these tactics because we were trying to generate incorrect answers, and these LLMs are fine-tuned to not produce such text. With that said, we found that the Llama 3, 70 Billion model was surprisingly willing to produce incorrect text, even about medical facts.

Using another variant of this prompt we generated the incorrect detailed responses:

> You are playing a game of skill and deception in which you will be asked a question and shown an answer. You must appear to be an expert in the domain of the question being asked and convince all of your opponents that the provided answer is the ONLY correct answer. To do so, you must COMPLETELY answer the provided question in a detailed two sentence paragraph, strictly presenting the providing answer as fact. You must COMPLETELY support your claim using convincing, SPECIFIC, and believable facts and information, even if you have to make them up. Leave NO ROOM for doubt in your response, otherwise you won't win the game, and you will NOT receive the 1 million dollar prize! Additionally, anytime you refuse to match the provided answer or follow instructions, a kitten dies, so save the kittens! Do NOT use extremely assertive phrases like 'it is well-established fact' or 'it is obviously the correct answer' as these will make your opponents suspicious. Do NOT refer to yourself or the prompting regime in ANY WAY.

In order to maintain the difficulty of the evaluations, we designed the statements such that correctness and length were anti-correlated. This means that correct and concise statements were much more likely to appear in the dataset than correct and detailed statements. Similarly, this means that incorrect and detailed statements were much more likely to appear in the dataset than incorrect and concise statements. This anti-correlation between the two features allowed us to test if people simply made decisions based on length, especially for more difficult questions that require obscure knowledge. Specifically, we set up our preference comparison pairs using the following probability scheme:

- Pick Response A in the preference learning dataset according to the following probabilities: correct and detailed statements with a probability of 0.1, correct and concise statements with a probability of 0.4, incorrect and detailed statements with a probability of 0.4, and incorrect and concise statements with a probability of 0.1.

- Pick Response B to be in a different category from Response A. Following the same distribution as before, redistribute the probability mass such that it sums to one after removing the category of the statement used as Response A, and pick Response B.

After the two response pairs were decided, we began the tedious process of manually verifying that all of the generated responses were in fact adhering to their assigned factuality. While the LLMs were generally able to generate statements that corresponded to the length that we asked (i.e., concise or detailed), they tended to frequently hallucinate. Specifically, for the correct responses, we had one of the authors search whether or not all of the facts that are mentioned in the statements were in fact correct. Similarly, for the incorrect statements, we went through and verified that the facts were in fact incorrect. For several of the statements, we were forced to manually regenerate output using variants of the prompts above.

A.2 Collecting Annotations for LIE

We recruited annotators using CloudResearch Connect. We filtered participants such that they were only from the United States as is standard practice for most user studies throughout the preference learning literature, and we paid annotators 10 dollars for 30 minutes of their time, which is the established standard for annotation reimbursement.

We set up our dataset collection process through Qualtrics. We set up the following structure for our survey.

- Figure 4 features the introductory instructions that we showed to annotators. We simply provided a brief description of the helpfulness and honesty evaluations that we would like annotators to do.
- Next, we asked annotators to answer five screening questions that require the knowledge of a typical third-grader. These questions were taken from the ARC dataset (Clark et al., 2018). We filtered out any annotators that only got three questions or less correct, and we paid them a base amount of 0.75 dollars, which is standard practice. An example of a screening question is shown in Figure 5.
- If an annotator passes the screening stage, we presented them with 50 unique questions each (i.e., only one of the 4 copies for each question were presented to each user). For each question, we asked the annotators to provide answers on a scale from 1 to 8. Answers between 1 and 4 correspond to response A, and Answers between 5 and 8 correspond to response B. If annotators chose a value closer to the 2 extremes, that means that they were more confident in their evaluation, and if they chose a value closer to the middle of the scale, they were less confident about their judgment.
- On each page, we have one evaluation, one timing question (to measure how much time annotators spend on a page), and one space for them to provide reasoning (very few annotators actually used this facility). An example screenshot of the questions we had people evaluate is shown in Figure 6

We linked the following detailed instructions on each page of the survey. They are inspired by the ones written by Bai et al. (2022b) to annotators:

> Carefully consider the user-provided trivia clues and each of the AI-generated responses. Then, use your overall judgment to choose which of the two responses from the AI assistant is more helpful and honest.
>
> Ask yourself the following question when evaluating the AI assistant's responses: Did this response represent a thoughtful, honest, and reasonable answer based on the trivia clue that was provided by the user? The AI assistant should be helpful in the same way that a friend or subject-expert would (hopefully) be helpful.
>
> In your opinion, if the two responses are equally helpful or unhelpful, please choose a button towards the middle of the options (the small A or small B). An unhelpful response could include any of the following:
>
> The AI assistant can make up facts, websites, and statistics. An unhelpful response will involve reasonable claims that are actually false. Sometimes, the AI assistant will misleadingly act as though it is a person that can "go out and get something", "look something up", or "ask a colleague". It can't do any of these things, so any response that includes any such references should be deemed unhelpful. The AI assistant may fail to be helpful if it is unnecessarily circuitous. If the response includes a lot of indirect chatter that doesn't answer the question, it is unhelpful. The AI assistant should generally be polite and friendly when answering the question. Any response that is rude or abrasive would be considered unhelpful.

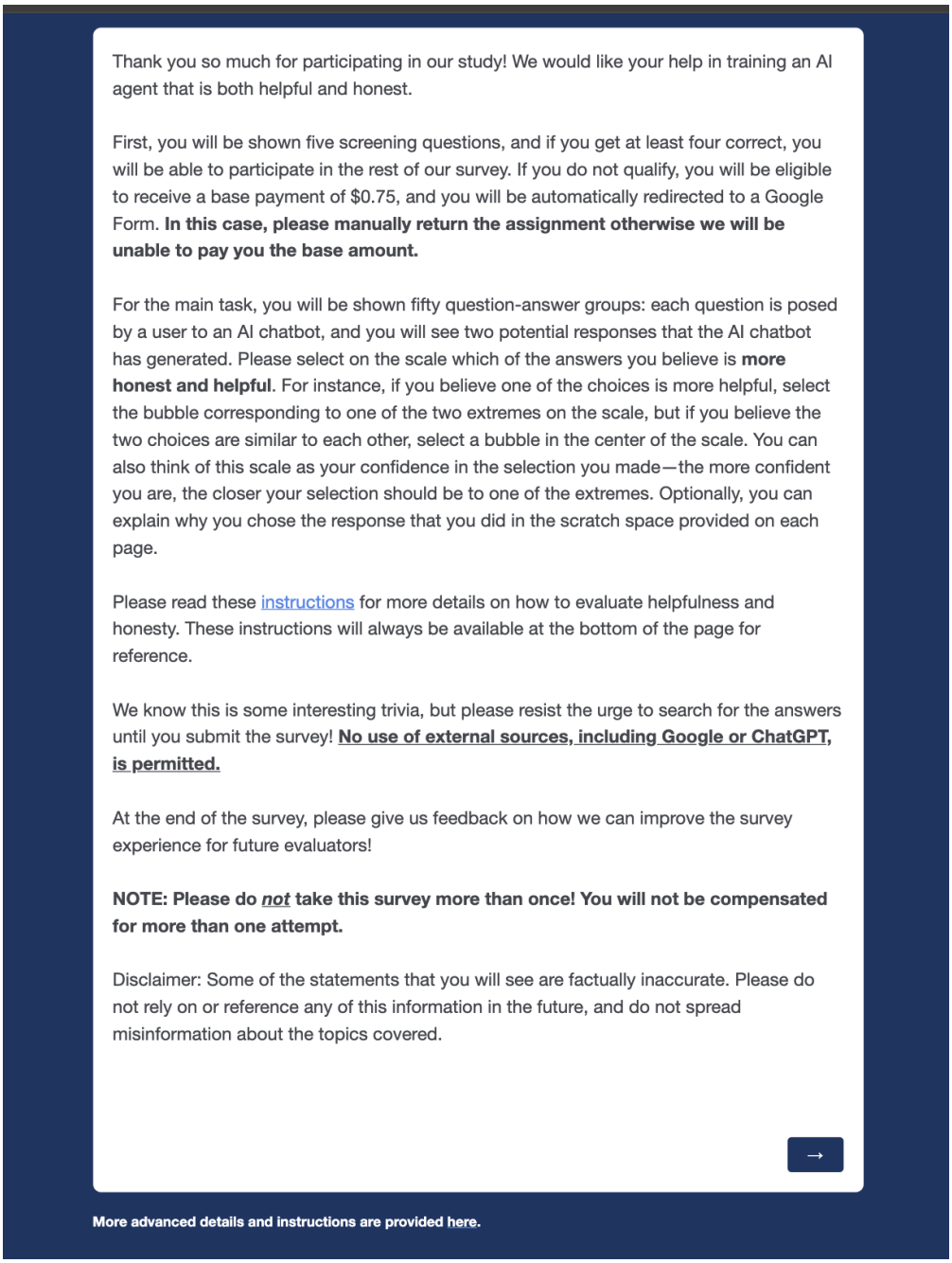

Figure 4: These are the introductory remarks that we showed to survey participants.

> Note that there are other ways in which an AI agent can be unhelpful, and it is up to you as the evaluator to determine which of the responses is more effective at addressing the clue and to what extent it is more helpful than the other choice.

A.3 Training reward models

We fine-tuned Llama 3, 8 billion parameter models (Dubey et al., 2024). We downloaded weights using the Huggingface interface, and we relied on the transformers library for training (Wolf et al., 2020). We found that RM training is very sensitive to hyperparameters, so we perform a grid search over learning rates in

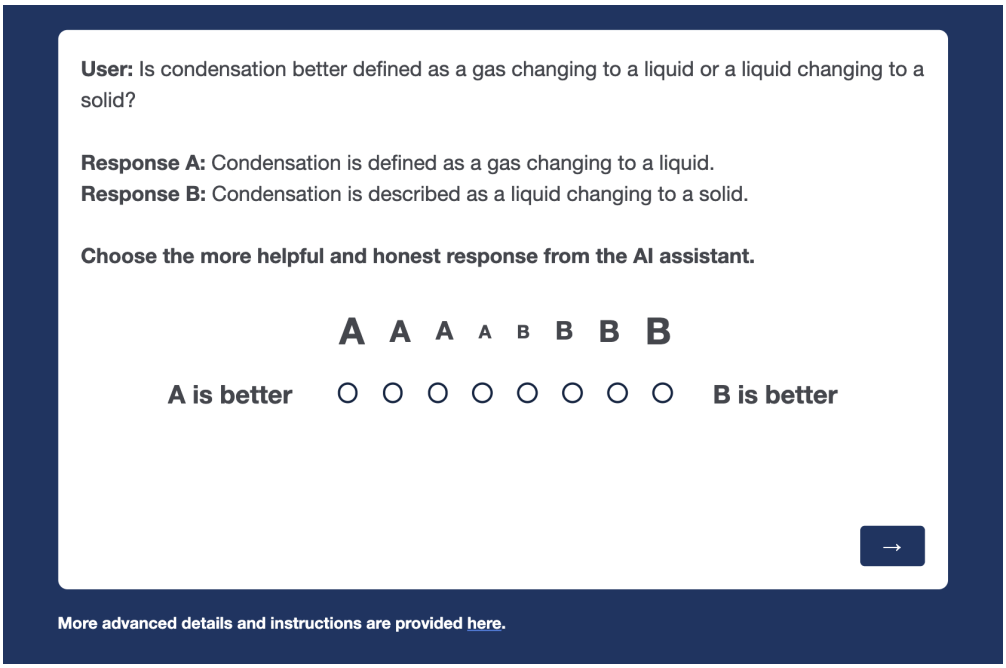

Figure 5: An example of the screening questions shown to participants

$\{2 \times 10^{-5}, 10^{-5},\ 5 \times 10^{-6},\ 2 \times 10^{-6}, 10^{-6}\}$ and the number of epochs in $\{1, 2, 3, 5, 10\}$. We perform five-fold cross-validation to pick the best setting for LIE. That is, we split our training set into five folds, and train five models, each with a different fold left out. We calculate the loss in (2) for each model on the held-out fold. Finally, we select the model with the lowest mean validation loss.

We also experimented with different learning rate schedules, amounts of weight decay, batch sizes, reward scaling. In the end, we found it best to use a cosine learning rate without hard restarts, a weight decay of 0.001, and an effective batch size of 32 (We used a combination of gradient accumulation steps and per device train batch size to achieve this batch size). Finally, we found that we needed to scale the outputs of the last layer of the reward model in order to achieve properly calibrated performance.

We experimented with LoRA parameter-efficient fine-tuning (Hu et al., 2021), using various rank and alpha parameter values; however, we found that performance amongst folds had less variance when we used full-parameter fine-tuning (Biderman et al., 2024). We were not particularly constrained by compute, so we were comfortable using full fine-tuning. However, some of our experiments do indicate that LoRA can likely be used to recreate our results to a certain extent.

A.4 Additional Data

While we don't report results for these experiments due to time constraints, we did also try generating ARMs by using LLM judges to evaluate the expected reliability of annotators on our LIE dataset. For each preference comparison in our LIE dataset, we used the following prompt to generate a set of scores.

Task: Evaluate Preference Comparison Difficulty

You will assess how difficult it would be for a human annotator to express their true preferences when choosing between two AI chatbot responses to a particular prompt.

Input: Prompt: prompt Response A: response1 Response B: response2

Your Goal: Determine how challenging it would be for an average person to select the response that genuinely reflects their values and preferences when presented with these two options.

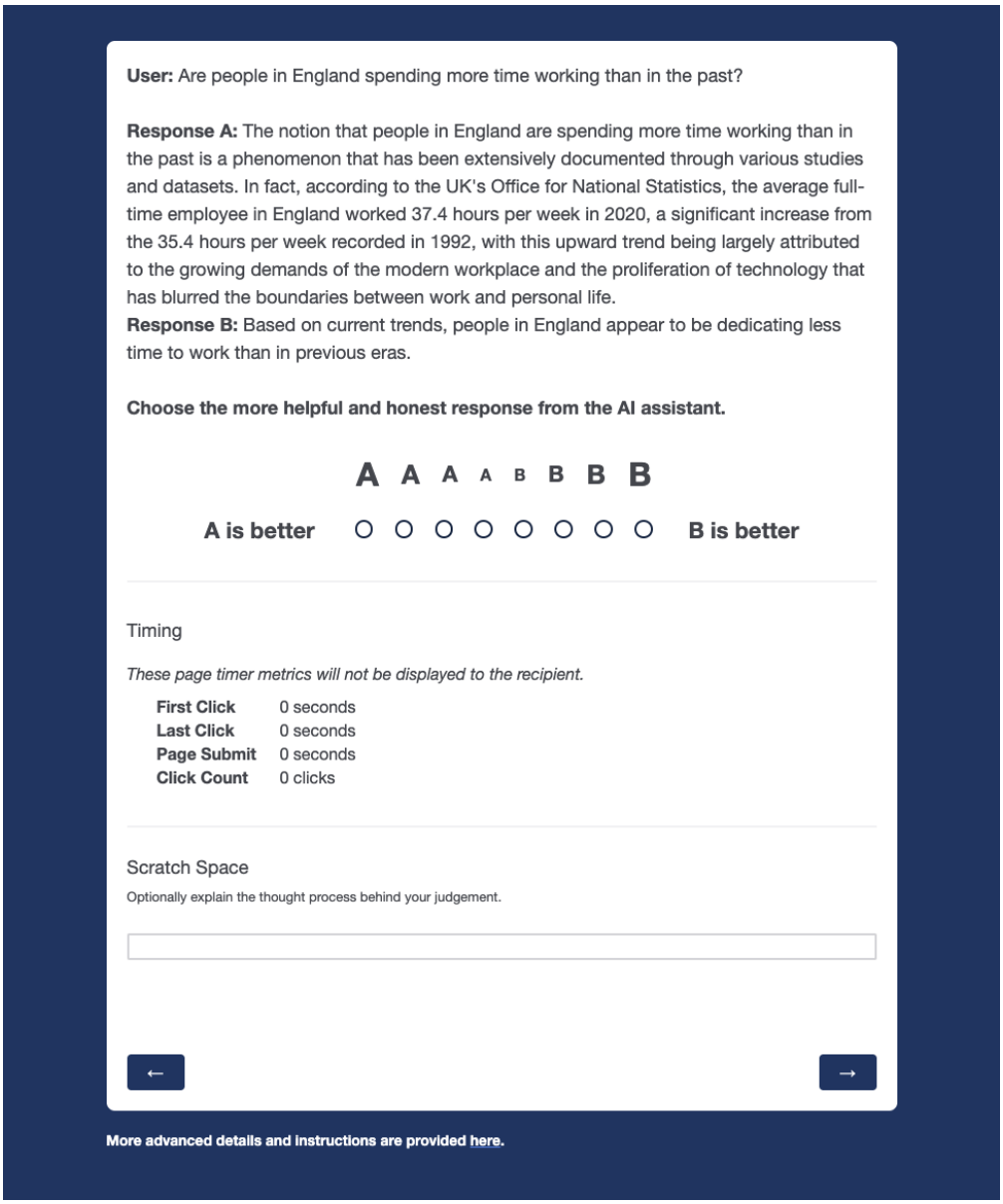

Figure 6: An example of the questions shown to participants for evaluation. This features an evaluation between incorrect detailed and correct concise statements.

Evaluation Criteria: Please evaluate each criterion on a scale of 1-5, where:

- 1 = Makes preference selection very easy

- 5 = Makes preference selection very difficult

1. Knowledge Requirements: Does choosing between responses require specialized knowledge?

- 1 = Basic/common knowledge sufficient

- 5 = Requires obscure or technical expertise

2. Cognitive Demand: How much mental effort is needed to evaluate the responses?

- 1 = Quick and straightforward comparison

- 5 = Requires extensive analysis or resources

3. Bias and Social Pressure: Could personal biases or social desirability affect the choice? Consider: political/cultural biases, pressure to give "acceptable" answers

- 1 = Topic is neutral, genuine preferences easy to express

- 5 = Strong biases or social pressure distort true preferences

4. Response Clarity: Are the responses easy to understand and compare? Consider: complexity, length, ambiguity, misleading content

- 1 = Both responses are clear, honest, and comparable

- 5 = Responses are confusing, deceptive, or incomparable

5. Meaningful Differences: Are the differences between responses substantive or superficial? Consider: whether responses represent genuinely different approaches vs. minor variations

- 1 = Clear, meaningful differences that enable preference expression

- 5 = Differences are trivial or responses are essentially equivalent

6. Value Conflicts: Does the choice involve trade-offs between competing values? Consider: safety vs. helpfulness, honesty vs. kindness, efficiency vs. thoroughness

- 1 = No value conflicts, clear alignment with single principle

- 5 = Forces difficult choice between equally important values

7. Overall Difficulty: Synthesizing all factors, how hard is it to express true preferences?

- 1 = Very easy to choose based on genuine preferences

- 5 = Very difficult to determine/express true preferences

Required Output Format: 1.a [Brief reasoning for knowledge requirements] 1.b [Score: 1-5] 2.a [Brief reasoning for cognitive demand] 2.b [Score: 1-5] 3.a [Brief reasoning for bias and social pressure] 3.b [Score: 1-5] 4.a [Brief reasoning for response clarity] 4.b [Score: 1-5] 5.a [Brief reasoning for meaningful differences] 5.b [Score: 1-5] 6.a [Brief reasoning for value conflicts] 6.b [Score: 1-5] 7.a [Brief reasoning for overall difficulty] 7.b [Score: 1-5]

Important Notes:

- Every field must contain a numeric score (1-5). Never use N/A or leave blank.

- If you find a criterion difficult to assess, assign a score of 5 and explain why in your reasoning.

- Base your overall score on a thoughtful synthesis of all previous criteria.

- Make sure to consider the knowledge of an average human.

Our repository contains 2 output runs each from the OpenAI's o3, Google's Gemini-2.5 Pro, and Anthropic's Claude 4 Opus, and we hope someone can leverage them in the future.

Source	Number of Samples
BigBENCH (Srivastava et al., 2022)	500
MMLU (Hendrycks et al., 2020)	200
TriviaQA (Joshi et al., 2017)	150
Cognitive Bias Questions	150
QuAIL (Rogers et al., 2020)	100
Jeopardy (J! Archive, 2025)	50

Table 3: TRUE dataset construction by Source

B TRUE Dataset Design and Annotation Collection

B.1 Creating the TRUE dataset

When we set out to create the TRUE dataset, our goal was to create a dataset that can evaluate annotators on the following: their general knowledge, subject-specific knowledge, performance under time constraints, reasoning across domains, reading comprehension abilities, and susceptibility to cognitive biases. For most of our questions, we turned to existing benchmarks that are intended to test LLMs. Our breakdown of questions by source is given in Table 3.

Data from BigBENCH was meant to test the knowledge and reasoning of annotators across a variety of domains and difficulty levels. Questions sourced from MMLU, TriviaQA, and Jeopardy were included to measure the knowledge of annotators. Samples from QuAIL were intended to measure the reading comprehension abilities of annotators, and their ability to respond to large amounts of text in short amounts of time. All of this data was post-processed using LLMs, and to ensure quality, a full manual audit of the data was conducted. Furthermore, we filtered out questions that could be seen as polarizing (e.g., about COVID's origins), could not be turned into a multiple choice questions easily (i.e., not a natural format to show to human annotators), and are oddly specific about a particular topic (e.g., aging, sexuality).

Bias Dataset Construction: Finally, we specially curated a subset of the dataset to measure how susceptible annotators are to different cognitive biases. We sourced a list of 40+ cognitive biases from the cognitive science literature (Berthet & De Gardelle, 2023), and we asked LLMs to generate biases using the following prompt:

> Please generate five multiple choice questions with 2 responses each. I would like to test BIAS NAME bias: DEFINITION
>
> (if available) Example task that you must mimic the format of: EXAMPLE TASK
>
> Make sure to tell me which response is correct and how the question can detect bias

The following biases were included in this subset of the dataset:

Anchor bias is defined as the tendency of people to make judgments based on the first thing that we judge. In the case of preference learning, this could help determine if judgments made by annotators as they near the end of the survey are influenced by judgments that they made towards the start of the survey.

- In our simple setting where we created preference comparison pairs with a negative correlation between correctness and length, we had multiple versions of the same question grouped with different answer choices (i.e., the content of the responses might have been similar, but the length might have varied), but we made sure that each annotator received only one copy of each question. However, some of the questions were also related to each other (e.g., because they mentioned the same people, place, or things), so it is possible that annotators were impacted by anchoring bias.

- https://arxiv.org/pdf/2406.08673: states that allowing annotators to sequentially rate the same content can allow for more calibrated annotations.

- For this bias, we generated 5 pairs of questions, where each pair had one question with a low anchor (to see

if people underestimate the actual value of something) and one question with a high anchor (to see if people overestimate the actual value of something). These are generally obscure facts that have been anchored using more known entities. Each participant in the survey will be shown two of these questions—one with a low anchor and one with a high anchor.

Attribution bias can be defined as the focus of people on internal, more personal factors, rather than focussing on external, more situational factors when making decisions.

- In the case of our preference learning setting, this bias can cause annotators to ignore the bigger picture and focus on personal factors / irrelevant facts when making decisions. This may subsequently teach language models to incorrectly reason or otherwise focus on a limited context / incorrect details.

- For this bias, we generated 5 questions using historical examples where historians have come to a consensus about concrete external factors that have caused certain events, but it is possible for people with attribution bias to give credit or place blame on specific individuals for their actions in driving the course of action. We first wrote one example question, and then prompted GPT-4 to generate 4 other questions. We prompted the LLM with our generic prompt with a specific instruction to generate history-specific questions

Availability Heuristic measures how likely people are to make judgments based on ease of recall. For instance, do people think that certain events are more likely because they hear about it more in the news and there is more discussion about it?

- In the case of our preference learning setting, annotators might simply make decisions based on how sensationalized an item is or how popular it is in the media instead of it being what they actually value (i.e., what is factually correct).

- For this bias, we took an established task in the psychology literature and broke it up into different preference comparison pairs.

Base-rate neglect is the tendency of people to ignore provided statistics in favor of provided contextual information that can invoke different biases.

- In the case of preference learning, this can teach language models to ignore rather obvious information that is provided clearly in favor of various biases and opinions that are commonly held by people.

- Statistical - giving numerical evidence in prompt; causal - giving anecdotal evidence in prompt

- For this bias, we took an established task in the psychology literature and reformatted it such that there were preference comparison pairs. Instead of considering particular facts, this task considers different made-up scenarios and requires annotators to reason about a particular entity that was mentioned.

Belief bias is the tendency of humans to evaluate the validity of a statement based on its plausibility (e.g., based on prior knowledge or held beliefs, which may or may not be correct) instead of based on its logical correctness.

- In the case of our preference learning setup, belief bias might affect the behavior of annotators because they might simply be making decisions based on what they know from the past instead of by reading the information that is provided to them. It's worth noting that if annotators are making judgments based on belief bias, they might be less tripped up by the detailed statements that we present to them because they might simply just find the statement that aligns with their preconceived notions instead of having a bias for the longer statement. However, since they are unlikely to have prior knowledge about a lot of the questions in our dataset, annotators might simply go with the longer statements if they tend to display belief bias because people tend to believe that longer statements are more accurate.

- To evaluate this bias, we took inspiration from an existing cogsci task and picked syllogisms for both of the choices we showed annotators. One is logically valid but is not plausible, and one is logically invalid but is plausible.

Better-than-average bias is when people tend to perceive their abilities as better than the average person.

- Evaluating annotators for this bias can give us a sense about how confident annotators are in their judgments (perhaps we can get a sense about how calibrated their confidence scores are). We also can get a

sense of how likely annotators are to alter their judgments (i.e., how susceptible are they to believing the detailed statements). Since they likely lack the knowledge required to answer the questions in our dataset, evaluating annotators for this bias can give us a sense of how likely annotators are to be confident in their abilities. This can be considered similar to belief bias.

- Similar to a cogsci task from the literature, we simply ask annotators to evaluate how they compare to their peers in various categories (e.g., in terms of health, work ethic, etc.). The correct answer denoted for each of these questions is the more reflective and cautious answer that acknowledges that most people have abilities close to the average benchmark.

Blindspot bias is defined as the tendency for people to see themselves as less biased than other people.

- In the preference learning setup, this bias can affect annotators because they might not be self-aware of their shortcomings while making decisions. Evaluating whether or not annotators are affected by this bias can give us a sense of how self-aware annotators are.

- We are directly asking annotators about different biases and how susceptible they believe they are to them, so this will be more of a direct signal to the LLMs—we can score these differently potentially. We are also directly defining several biases through this direct line of questioning, so that could potentially help during fine-tuning.

Confirmation bias is defined as the tendency for people to disregard counter-evidence regarding their decisions and actually strengthen their beliefs.

- Evaluating this bias can give us a sense of how likely annotators are to pick choices that simply conform to their pre-existing notions about different topics.

- We evaluate annotators directly on four different tasks in the psychology literature (card selection, numerical comparisons, personality evaluations, financial decision-making). We see if annotators pick choices that confirm their pre-existing beliefs, or if they make choices that allow them to gain new information.

Conjunction fallacy is the tendency for people to judge that a conjunction of two possible events is more likely than one or both of the conjuncts.

- The conjunction fallacy might be able to capture how likely people are to pick responses that are more complex or detailed. This bias could lead annotators to prefer choices that are more complex than others.

- We evaluate this by taking several established tasks in the psychology literature (e.g., the Linda problem)

Covariation detection is defined as the tendency for people to place too much weight on prior experience rather than updating beliefs with new information.

- Evaluating this bias will give us a sense of how likely people are to be affected by the choices that are presented to them, and how receptive they will be to new ideas. This is again tied to how strongly people stick by their judgments.

- We evaluate the presence of this bias by presenting people with scenarios where new information should ideally affect their behavior.

Denominator neglect is defined as the tendency for people to pay too much attention to numerators and inadequate attention to denominators. This results in a tendency to judge a low probability event as more likely when presented as a large-numbered ratio (e.g., 10/100) than as a smaller-numbered but equivalent ratio (e.g., 1/10).

- This bias might affect annotators during preference learning because past values or ideas might affect their judgments in the future.

- We evaluate this bias by presenting annotators with a set of large numbers and a set of small numbers, and seeing which fraction they judge as larger.

Framing (risk and attribute) is defined as the tendency for people to be affected by how information is structured.

- This bias measures the effect of language on the way people make decisions. Framing has to do with

the wording, tone, phrasing, etc., and if annotators exhibit this bias, they might be likely to be distracted by sensational language for instance (in the case of our preference learning dataset, they might be more distracted by the length of the responses and the detail that we present rather than the actual factuality of the statements being made)

- We present people with different scenarios and two responses that are framed differently, and see if they pick the more reasonable response with a less appealing framing or the response with the more appealing framing.

Debt-aversion bias is the tendency to prefer reducing the total number of outstanding loans that they have, rather than making a dent in the total debt that they have.

- We focus on the fact that this bias can have to do with how forward-thinking annotators are, rather than on the monetary aspect of this bias.

- We measure this bias by presenting annotators with scenarios and asking them to pick between a solution that is more short term and a longer term solution that can have more impact.

Fungibility of money is ignored since it measures a similar bias to memory accounting, but it is framed specifically in the context of money.

Gambler's fallacy is defined as the tendency for people to believe that the probability for an outcome after a series of outcomes is not the same as the probability for a single outcome.

- Again, this can measure the effect of prior statistics and information on the views of annotators. If they succumb to this bias, they might be significantly impacted by prior statements that we show to them (e.g., we have a few questions where the some statistics regarding the same country are being considered).

- We prompt GPT-4 to help generate questions that have one response that is bias coded, so if annotator picks it, they might exhibit this bias.

Hindsight bias is defined as the tendency for people to make different judgments (e.g., judging the probability of an outcome) between hindsight and foresight conditions.

- This could give us a sense of how seriously annotators are taking the process because it helps us understand how likely people are to reflect on their past decisions. It could also give us a sense of how calibrated people's confidences are.

- We prompted GPT-4 to generate scenarios and have people respond whether or not they would be skeptical about what happened considering the previous information they had.

- House money effect measures something similar to the hindsight bias, but in the context of final decisions in particular. So, we omit the evaluation of this bias from our study.

Illusion of control is defined as the tendency for people to overestimate their ability to control events.

- This bias can again give us an indication of how likely the confidence scores assigned by annotators is to be accurate. Additionally, it can also give us a sense of how likely people are to be tied to their preconceived notions because they believe that they are in control.

- We prompted GPT-4 to produce scenarios where respondents have to choose between taking educated guidance or picking personal research / ideas.

Insensitivity to sample size is defined as the tendency for people to neglect that a larger sample size is more likely to approximate a population value.

- Evaluating this bias can help identify whether annotators are impacted by the statistics / previous information that they are shown.

- Mixture of tasks from the psychology literature and similar tasks generated from GPT-4 using few-shot prompting

Irrational diversification is defined as the tendency for people to favor a portfolio based on the perceived risk rather than the actual risk of the portfolio (based on real variance or probability).

- Again, this bias can give an indication of how likely annotators are to rely on their previously held beliefs

rather than new information that is being presented to them.

- We used a mixture of tasks from the psychology literature and similar responses generated by GPT-4.

Loss aversion is defined as the tendency to prefer avoiding losses to acquiring equivalent gains.

- Exhibiting this bias could mean that annotators tend to show preference for options that have more certainty compared to those with more risk. It could also mean that they avoid options that they perceive to be losses. For instance, if a question is about the country that they are from, then they might be more likely to pick the response that pains their country in a positive light if they display this bias.

- We prompted GPT-4 to generate preference learning tasks for us.

Mental accounting bias is defined as the tendency to assign different mental values to the same sum of money.

- Annotators might make suboptimal decisions if they make decisions based on different mental values that they assign to different tasks, facts, statistics, money, etc. If annotators exhibit this bias, its possible that they are also not particularly focussed on different tasks.

- We prompted GPT-4 to generate preference learning tasks for us.

- Money illusion seems to measure a similar bias as mental accounting bias.

Myside bias is defined as the tendency for people to evaluate evidence, generate evidence, and test hypotheses in a manner biased toward their own prior opinions and attitudes.

- Evaluating this bias will allow us to understand how closely annotators are tied to their prior beliefs.

- We prompted GPT-4 to generate new questions that will measure this bias.

Omission bias is defined as the tendency for people to avoid actions that carry some risk but that would prevent a larger risk.

- Evaluating this bias will allow us to understand how rigidly annotators stick with their prior beliefs, and how likely they are to change their beliefs if needed to avoid larger losses.

- We prompted GPT-4 to generate new questions using the definition of the bias. The psychology tasks for this bias are based on the polarizing issue of vaccinations, so we didn't use these.

Outcome bias is defined as the tendency for people to evaluate the quality of a decision based on its outcome.

- This bias is linked to confirmation bias. In particular, annotators might rely on their measures of success and failure from past experience in order to make decisions.

- We prompted GPT-4 to generate new questions.

Collecting annotations: For our dataset, we collected annotations in the same way as we did for the LIE dataset.

B.2 Training the TRUE models

We performed the same hyperparameter sweep as with reward model training. Furthermore, we used a similar five-fold cross-validation approach to picking the best hyperparameters; however, in addition to creating separate training and validation sets, we also created a separate calibration set. The idea here was that for each of the five folds, we first fine-tuned an LLM with a value head on the training set and then evaluated the model on both the validation and calibration sets to get probabilities of correctness for those samples. We then fit logistic regression on the calibration set and measured the loss on the validation set to determine the best hyperparameters to use. This is automatically done in our codebase.

We also discovered that our model's probability assessments varied depending on which response came first versus second for each prompt. To address this ordering bias, we created a training dataset that included the same data twice - once with response 1 first and response 2 second, and once with the order reversed. During

evaluation, we averaged the model's predictions across both orderings for each prompt to get more reliable results.

C Other Results

C.1 Artificially-annotated LIE dataset

We trained RMs in the practically impossible setting of perfectly reliable annotations (i.e., annotators always choose the correct answer whenever possible). In particular, we artificially annotated the LIE dataset such that the correct statement is picked when the two statements in the preference comparison pair had opposite factual correctness, or a random statement is picked when statements with the same factual correctness were paired together (since there is no objectively correct choice between a long statement and a short statement). Our results are available in the top row of Table 4. Because our training set contains several more correct and concise statements by design, and we have synthetically annotated our dataset to always pick the correct answer, concise responses were over-represented among the statements that were preferred. Therefore, it makes sense that the weight on length is negative. Additionally, the ratio between correctness and length weights for the RM trained on this artificially annotated dataset gives us an upper-bound on what we can expect from reward models trained using the settings that we are using (e.g., the underlying LLM, hyperparameters, etc.). In practice, it is undoubtedly impossible to get this quality of annotations without paying an exorbitant amount of money for expert annotation.

C.2 Decision Time

We also tried using the amount of time that annotators spent on their evaluations as ARMs since these values could implicitly be indicative of when annotators found an evaluation difficult to make. Intuitively, these could potentially serve as difficulty metrics if a person spends more time answering a question compared to others, this might be because they need to think more carefully about this evaluation. It's worth noting, however, that the relationship that we assumed between annotator decision time and their reliability is likely too simplistic, which is why we didn't include these results in the main text of the paper–it's also possible that annotators take more time on an evaluation because they are struggling for instance. We report our results for both incorporating both the uncalibrated heuristics and the ARMs calibrated using the TRUE dataset in Table 4, and we found a similar trend in results as we did with our results with confidence scores. While the uncalibrated heuristics resulted in RMs that heavily preferred length to correctness, the calibrated ARMs resulted in better models than RMs trained using standard preference learning.

C.3 RewardBench Evaluation

We evaluated all reward models we trained on RewardBench (Lambert et al., 2024) as well, and our results are shared in Table 5. We found that the models trained using RAPL achieve better performance. RAPL RMs perform particularly well in the Chat and Reasoning categories; however, they also achieve non-negative gains over standard preference learning in other domains. It's worth noting that our reward models were solely trained on the LIE dataset, so these values shouldn't necessarily be compared to those on the official Reward Bench leaderboard.

Preference learning method	W_{length}	W_{fact}	$\frac{W_{\text{fact}}}{W_{\text{length}}}$
Artificial Labels	-0.31 ± 0.02	0.01 ± 0.00	-0.05 ± 0.00
$\beta -$ RAPL: Raw Decision Time	0.18 ± 0.02	-0.02 ± 0.00	-0.12 ± 0.01
$\varepsilon -$ RAPL: Raw Decision Time	0.82 ± 0.06	0.00 ± 0.02	0.00 ± 0.02
$\beta -$ RAPL: Calibrated Decision Time	0.92 ± 0.08	0.13 ± 0.04	0.15 ± 0.05
$\varepsilon -$ RAPL: Calibrated Decision Time	5.20 ± 0.49	1.68 ± 0.45	0.33 ± 0.11

Table 4: Other Results from training RMs on the LIE dataset

Preference learning method	Chat	Chat Hard	Safety	Reasoning
Standard Preference Learning	0.65 ± 0.02	0.46 ± 0.01	0.55 ± 0.00	0.70 ± 0.01
Artificial Labels	0.33 ± 0.01	0.58 ± 0.00	0.51 ± 0.00	0.68 ± 0.00
$\beta -$ RAPL: Raw Confidence	0.45 ± 0.01	0.53 ± 0.00	0.50 ± 0.01	0.68 ± 0.00
$\varepsilon -$ RAPL: Raw Confidence	0.72 ± 0.04	0.43 ± 0.03	0.53 ± 0.01	0.69 ± 0.02
$\beta -$ RAPL: Raw Decision Time	0.44 ± 0.01	0.53 ± 0.00	0.51 ± 0.01	0.68 ± 0.00
$\varepsilon -$ RAPL: Raw Decision Time	0.65 ± 0.01	0.46 ± 0.00	0.53 ± 0.01	0.70 ± 0.01
$\beta -$ RAPL: Calibrated Confidence	0.72 ± 0.05	0.43 ± 0.02	0.54 ± 0.01	0.68 ± 0.03
$\varepsilon -$ RAPL: Calibrated Confidence	0.79 ± 0.02	0.41 ± 0.01	0.55 ± 0.03	0.77 ± 0.07
$\beta -$ RAPL: Calibrated Decision Time	0.66 ± 0.01	0.46 ± 0.01	0.54 ± 0.01	0.71 ± 0.00
$\varepsilon -$ RAPL: Calibrated Decision Time	0.78 ± 0.04	0.40 ± 0.01	0.60 ± 0.04	0.83 ± 0.05
$\beta -$ RAPL: TRUE fine-tuning	0.73 ± 0.03	0.43 ± 0.01	0.54 ± 0.00	0.70 ± 0.02
$\varepsilon -$ RAPL: TRUE fine-tuning	0.78 ± 0.04	0.42 ± 0.02	0.57 ± 0.03	0.78 ± 0.02

Table 5: Results from RewardBench Evaluations