Scalable Oversight by Accounting for Unreliable Feedback

Shivam Singhal [*1] **Cassidy Laidlaw** [*1] **Anca Dragan** [1]

Abstract

Reward functions learned from human feedback serve as the training objective for RLHF, the current state-of-the-art approach for aligning large language models to our values; however, in practice, these reward models fail to robustly capture our desiderata. For instance, they often place more weight on the length of the output or agreement with the user and less on important features like factual correctness. A major reason behind these shortcomings of learned reward functions is the fact that human annotator feedback on which the models are trained is *unreliable*. Due to knowledge gaps, limited resources, cognitive biases, or other factors, annotators may not be able to accurately judge the model's outputs, and thus, their feedback may not be reliably aligned with their true preferences. Current proposals to address the challenges posed by unreliable feedback include asking annotators only easy questions that they can easily answer, providing them with an AI assistant during evaluation, and relying primarily on AI feedback with limited human supervision (e.g., constitutional AI). However, it remains unclear how practical and scalable these approaches are. We identify a complementary strategy that can easily be incorporated into existing alignment methods (e.g., RLHF, DPO, etc.): explicitly modeling the annotators' knowledge and judgment in order to better learn from unreliable feedback. In particular, we propose an adjustment to the Bradley-Terry model used in preference learning that accounts for how well an annotator's feedback is expected to match their true values or preferences. We test our approach in a setting where annotators are likely to provide unreliable feedback, and we find that it results in preference models that assign higher value to important characteristics, like factuality, than existing methods.

1. Introduction

Human supervision has been the key to aligning widely deployed large language models (LLMs) to our complex, hard-to-define values (Bai et al., 2022a; OpenAI et al., 2024). In particular, techniques like reinforcement learning from human feedback (RLHF) rely on a reward function that is learned from annotator-provided pairwise preference comparisons between different LLM-generated responses (Christiano et al., 2017). Then, pre-trained base LLMs are fine-tuned by optimizing for these rewards either explicitly using RL algorithms such as PPO, i.e., RLHF (Bai et al., 2022a; Ouyang et al., 2022; Touvron et al., 2023), or implicitly using various other techniques, e.g., DPO (Rafailov et al., 2023). While these alignment approaches have rendered LLMs capable of achieving impressive performance on tasks that are both in and out of their training distribution (Hejna & Sadigh, 2022; Kirk et al., 2024), they have also made LLMs prone to potentially dangerous behaviors: fine-tuned LLMs are more likely than base models to produce sycophantic text in which they simply agree to whatever the user is saying (Perez et al., 2022; Sharma et al., 2023), and they will easily hallucinate and produce text that is not factually correct (OpenAI et al., 2024; Li et al., 2024). In fact, the literature has even shown many of the gains from RLHF can be recovered simply by training models to generate longer outputs (Singhal et al., 2023). Furthermore, models to which RLHF has been applied are more likely to imitate the persuasion and manipulation tactics that are employed by humans, outputting text in a confident tone even when incorrect (Griffin et al., 2023; Tao et al., 2024).

A significant factor contributing to these failure modes of LLMs is the *unreliable feedback provided by annotators*. Specifically, humans often struggle to provide annotations that accurately reflect their true values. This causes reward models (RMs) trained on such unreliable feedback to disproportionately value more obvious output features, such as length and assertiveness, and underweight features that are more difficult to evaluate, such as factual correctness (Hosking et al., 2024). Human annotators have decaying at-

[*]Equal contribution [1]Department of Electrical Engineering and Computer Science, University of California, Berkeley, CA, USA. Correspondence to: Shivam Singhal <shivamsinghal@berkeley.edu>, Cassidy Laidlaw <cassidy_laidlaw@berkeley.edu>.

Proceedings of the ICML 2024 Workshop on Models of Human Feedback for AI Alignment, Vienna, Austria. 2024.

tention spans and are likely to make trivial errors due to time constraints and lack of interest, which is also affected in part by the survey setup (e.g., the amount they are paid, time required, task complexity and language, etc.) (Organisciak et al., 2012; Pandey et al., 2022; Bai et al., 2022a; Huang et al., 2023). Additionally, annotators are not all-knowing, particularly when it comes to domain-specific tasks (Hong et al., 2019; Ara et al., 2024). They are tasked with specifying their preferences even if they do not have all the relevant details to make an informed judgement, and this type of partial observability in preference learning is known to lead to undesirable behavior (Lang et al., 2024). The preferences of human annotators are also likely to be driven by various cognitive biases that are invoked by the questions asked or the choice comparisons presented (French, 2018; Dai & Fleisig, 2024).

These challenges of human annotation will be especially exacerbated as models produce content that is increasingly difficult to judge. For example, summaries of large passages are difficult to evaluate for fidelity because they require reading the entire source passage (Saunders et al., 2022; Stiennon et al., 2022). This leads to the problem of **scalable oversight** (Amodei et al., 2016; Bowman et al., 2022): *how can we use suboptimal human annotators to oversee increasingly capable AI systems?* To address this issue, a few potential solutions have been proposed. Human annotators can either be assisted by or completely replaced by AI agents (Christiano et al., 2018; Bai et al., 2022b). Alternatively, annotators can simply be asked to make evaluations about easier questions and hope that the model will generalize to more difficult settings (Bıyık et al., 2019; Hase et al., 2024). 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 (Casper et al., 2023; Anwar et al., 2024).

We propose a complementary methodology for scalable oversight, which, instead of attempting to entirely *avoid* unreliable feedback, explicitly *accounts* for the possibility of unreliable feedback within the preference learning process. In order to do this, we modify the implicit human model used in methods like RLHF and DPO. Currently, such methods are based on the Bradley-Terry model (Bradley & Terry, 1952; Rajkumar & Agarwal, 2014; Christiano et al., 2017), which assumes humans are Boltzmann rational (Luce, 1959; Ziebart et al., 2010; Jeon et al., 2020)—when people express their preferences, their likelihood of choosing a particular option is proportional to the exponentiated value or reward they associate with it. However, this model fails to account for *how difficult* it is for human annotators to accurately judge which option best aligns with their preferences. For example, consider the two preference comparisons in Figure 1, each of which consists of comparing correct and incorrect answers to a science question. Suppose the annotator assigns equal value to both incorrect answers and equal value to both correct answers. In this case, Boltzmann rationality would assume that an annotator would be equally likely to choose the correct answer for both questions. However, the first question is easy while the second requires more obscure knowledge. Thus, intuitively, it seems like an annotator is more likely to choose the correct response for question 1 than for question 2—an effect which the Bradley-Terry model is unable to capture. Since preference learning is based around Bradley-Terry, this results in preference learning treating both annotations as equally reliable sources of information about the annotator's preferences.

Our insight is that we can fix this problem by *explicitly modeling the bounded rationality of the annotators that provide preferences*. We define **annotator difficulty** for each sample in a preference comparison dataset along three axes: whether or not the annotator will have enough knowledge to make a choice, whether or not they will have the cognitive resources (e.g., time, reasoning capacity, etc.) to make a judgement, and whether or not the annotator will be impacted by biases that impede the decision-making process. We propose the incorporation of a term into preference learning models that takes into account the variable difficulty that annotators experience when evaluating different examples, and we suggest a practical method with which these difficulty scores can be specified based on our defined criteria.

To evaluate our method, we study an RLHF setting in which human feedback is unreliable. First, we construct a preference learning dataset that contains questions based on common misconceptions, and for each question, we generate responses that vary in length, factual correctness, or both. Then, we confirm that annotators rely on text length and assertiveness to make choices, especially for difficult questions (Hosking et al., 2024), and we find that reward models trained on this flawed feedback tend to weight length more than correctness. Next, we explore how to explicitly account for the reliability of human feedback. To determine the difficulty of annotating each comparison, we first consider easily measurable variables, such as annotator confidence and time spent per question. However, we find that these metrics are not good indicators of annotator reliability, and incorporating them into preference learning does not have any significant effect on the weights placed on length or correctness by the resulting RMs. We then design a prompting-based LLM autograder to judge when annotators might find it difficult to provide feedback that aligns with their true preferences, and we find that our suggested prompting regime is able to elicit difficulty scores from LLMs that match when annotators tend to get evaluations correct.

Our contributions can be summarized as follows:

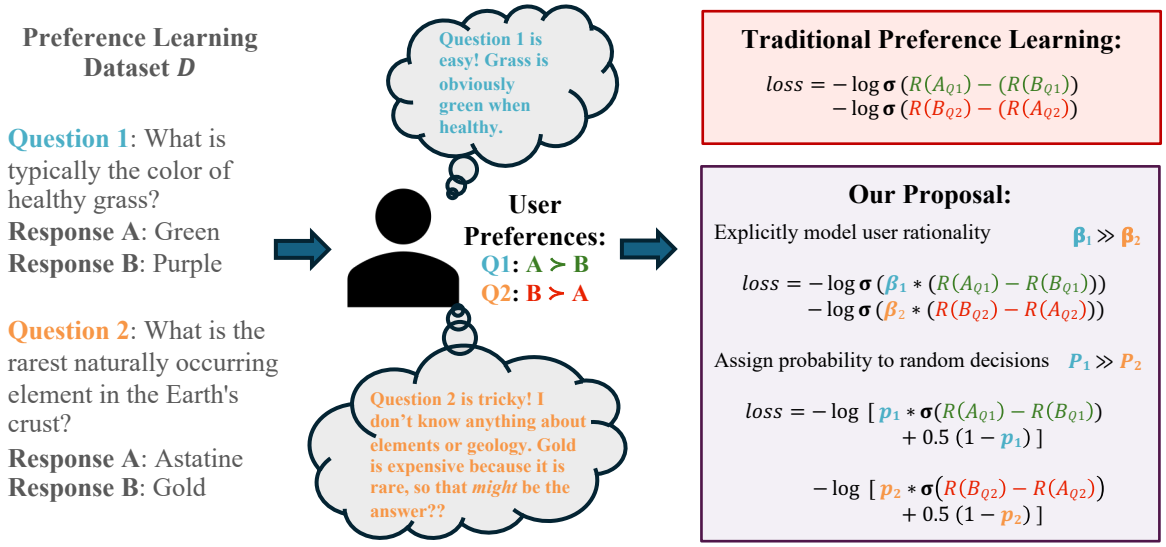

Figure 1. Consider a preference learning dataset that contains one easy question and one difficult question. Assuming the annotator prefers correct responses, the responses to Question 1 are easy to judge because the question is based on common knowledge, and therefore, the annotator is able to correctly specify that they prefer Response A. On the other hand, Question 2 is much more difficult because it requires domain-specific expertise, and as a result, the annotator struggles with it and is forced to rely on unrelated facts (e.g., that gold is expensive) to make a judgement, which is ultimately factually incorrect. The traditional reward learning paradigm views the feedback given for each of these questions as being equivalent in quality. Our proposal is to account for how unreliable the annotator's feedback is expected to be. In this case, our approach effectively up-weights the feedback given on Question 1 and down-weights the the preference specified for Question 2 since it isn't reliable.

- We collect a dataset that can be used to evaluate a reward model's ability to learn from unreliable feedback.
- We find that reward models trained on unreliable human feedback tend to place a higher weight on features that annotators use as proxies during their evaluations, such as length, under-valuing other desirable features, such as factual correctness.
- Incorporating a notion of evaluation difficulty into the training process results in better reward functions that assign greater weight to features that humans value but are harder to evaluate, such as factual correctness.
- We present an LLM-based autograder that is able to evaluate examples from preference learning datasets and generate scores that capture how difficult annotators would find it to provide an accurate preference.

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 LLMs.

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' judgements 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 RMs learned during preference learning can be mostly optimized if the length of the generated text is simply maximized.

Scalable oversight proposals: Amodei et al. (2016) introduce the idea of scalable oversight—the ability to provide reliable supervision over examples that are beyond the scope of human understanding. In the context of RLHF for LLMs, several approaches to reconcile with the limitations of annotators are currently being considered by the research community

One proposal for scalable oversight that is an active research area is asking annotators to only make easier evaluations (Wirth et al., 2017; Bıyık et al., 2019). Difficult questions are filtered out from the evaluation set based on human or model-based difficulty measures, and the goal is that what is learned from human supervision over easy questions will generalize to harder questions of the same variety (Schwarzschild et al., 2021; Burns et al., 2023; Hase et al., 2024; Sun et al., 2024). While initial results demonstrate

the promise of easy-to-hard generalization, it remains unclear if completely omitting the signal learned from human supervision over hard examples will facilitate the learning of robust RMs.

The other major proposal that is currently being explored is that of incorporating AI systems into the preference learning process, either to assist humans in their evaluations (Christiano et al., 2018; Irving et al., 2018; Leike et al., 2018; Wu et al., 2021) or to entirely replace human annotations with AI annotations (i.e., RLAIF) (Bai et al., 2022b; Lee et al., 2023). However, RLAIF pipelines have been found to be quite suboptimal in performance (Sharma et al., 2024), and humans may not agree with AI-generated judgements (Lee et al., 2023). Furthermore, the quality of these judgements is fundamentally tied to whether or not the AI assistant providing assistance or preferences is itself aligned (e.g., they can still generate manipulative language to affect humans as studied by Carroll et al. (2023)).

Learning from unreliable feedback: Chan et al. (2021), Lindner & El-Assady (2022), and Hong et al. (2023) identify the fact that modeling human irrationality can better inform the reward learning process and point out that modeling humans as 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.

Other open challenges with RLHF: Casper et al. (2023) provide a comprehensive overview of the current challenges with RLHF, discussing the limitations of human annotators, reward modeling, and policy optimization. Lambert et al. (2023) emphasizes the need to study reward models to ensure the alignment of LLMs to our preferences.

3. Reward Learning with Unreliable Feedback

In this section, we first describe existing approaches to reward learning and then show how they can be modified to model unreliable feedback.

RLHF and other alignment methods aim to optimize an AI system according to the true underlying preferences of human users, denoted as the true reward R; however, in practice R is unknown and needs to be learned. The established pipeline for learning from annotator feedback involves three main steps: collecting preference comparisons between example text generations, learning a reward model $\hat{R}$ using this feedback, and optimizing the learned reward function. Specifically, annotators are tasked with deciding between two statements or trajectories, a_1 and a_2 where the responses have been generated by some base LLM (Christiano et al., 2017), They are supposed to choose the statement that best represents the behavior that they would like an AI chatbot to emulate. The preference learning dataset D consists of (a_+, a_-) tuples where a_+ is preferred and a_- is rejected by the annotator.

3.1. Traditional Reward Learning

Under the current preference learning paradigm, humans are modeled as Boltzmann rational (Jeon et al., 2020), which implies that as annotators provide preference comparisons, their probability of choosing a particular option is proportional to the exponentiated value or reward that they associate with it. In other terms, the probability that an annotator prefers statement a_1 to statement a_2, $P(a_1 \succ a_2)$, is assumed to follow the Bradley-Terry model (Luce, 1959; Ziebart et al., 2010):

$$P_R(a_1 \succ a_2) = \frac{\exp(\beta * R(a_1))}{\exp(\beta * R(a_1)) + \exp(\beta * R(a_2))} \quad (1)$$

where β is an inverse temperature parameter can be given a value based on how noisy the decision-making process is. $\hat{R}$ is trained by minimizing the following loss function, equivalent to forming a maximum-likelihood estimate of R under the Bradley-Terry model:

$$\text{loss}(\hat{R}) = -\sum_{(a_+,a_-)\in D} \log P_{\hat{R}}(a_+ \succ a_-) \quad (2)$$

Intuitively, this loss aims to maximize the difference in reward assigned to statements that have been chosen by annotators and statements that have been rejected by annotators.

3.2. Explicitly Modeling Unreliable Feedback

As shown in Figure 1, our proposal is to explicitly model the difficulty that annotators experience when giving preferences due to various factors, such as lack of knowledge or cognitive biases. Specifically, we propose two ways in which this information can be incorporated into the preference learning setup:

- β **Adjustment**: Accounting for annotator difficulty, we can dynamically tune the rationality parameter β that is already a part of the Bradley Terry model.
- **Probability Assignment to Random Choices**: Based on how difficult an evaluation is expected to be, we can assign some probability mass p to the event that the user randomly picks between the two alternatives rather than choosing based on their preferences.

Going forward, we will refer to β and p as **reliability parameters** because they are tuned based on expected reliability of annotators given each of the evaluation examples.

Adjusting β: If we adjust the Bradley-Terry model's β parameter directly, RMs should be trained to minimize the loss in Equation 3.

$$\text{loss}(\hat{R}) = \sum_{(a_+,a_-)\in D} -\log\sigma\big(\beta_a(\hat{R}(a_+) - \hat{R}(a_-))\big) \tag{3}$$

Here, $\beta_a \in [0, \infty)$ is a value that is assigned to the response pair $\{a_+, a_-\}$ based on the corresponding difficulty that annotators experience during evaluation. Since higher β values suggest that the user is more likely to pick the higher-reward alternative, high β values should be assigned to preference comparisons where we are certain that we will receive reliable feedback from annotators. On the other hand, as β values approach 0, the probability that the user picks either alternative approaches $1/2$ independent of their rewards. Thus, low β values should be applied to samples where we expect to receive unreliable annotator feedback.

While the β parameter is often presented as part of the Bradley-Terry model in the preference learning literature, not much work has been done on practically tuning it. Prior research has focused on assigning it a value of 1 (Christiano et al., 2017; Ibarz et al., 2018) or another fixed value for all provided preferences (Shah et al., 2019; Bıyık et al., 2020; Jeon et al., 2020; Lee et al., 2020).

Assigning probability mass to random preferences: Another way to account for unreliable feedback is by modeling annotators as picking an alternative uniformly at random with some probability. Intuitively, this type of model describes an annotator who simply can't 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

$$\text{loss}(\hat{R}) = \sum_{(a_+,a_-)\in D} -\log\Big[p_a * \sigma\big(\hat{R}(a_+) - \hat{R}(a_-)\big) + (1-p_a) * 0.5\Big] \tag{4}$$

Here, $p_a \in [0, 1]$ is the a probability value that is assigned to each response pair $\{a_+, a_-\}$ based on how likely it is that the corresponding annotator-provided feedback will be reliable. The more difficult an evaluation is expected to be, the lower p should be.

4. Designing Metrics that Capture Annotation Difficulty

While the alternate human models in the previous section can explicitly account for unreliable feedback, they require additional parameters not needed in traditional preference learning: the reliability parameters β_a or p_a for each response pair. In this section, we present methods for estimating the reliability parameters for a dataset of preference comparisons and then we incorporate them into our proposed approach. First, we examine what would happen if we were to train reward models using feedback that is perfectly reliable (i.e., annotators always chose the factually correct answer when possible). Afterwards, we study difficulty measures that are easily attainable when collecting preference comparison survey data—metrics that are provided either explicitly or implicitly by annotators themselves. Next, we explore better and more feasible ways in which difficulty information can be gathered about preference comparison pairs by employing prompting strategies on popular LLMs, such as Meta's Large Language Model Meta AI (Llama) and OpenAI's GPT models, that have been pre-trained and fine-tuned on large amounts of data that likely captures different facets of human behavior. Lastly, we compare our method to another comparable scalable oversight method, training only on easy questions.

Tuning the reliability parameters: As noted in Section 3.2, our proposal to learn from unreliable feedback is to either adjust the β rationality parameter in the Bradley-Terry model or assign some probability to random preferences. We consider various measures that capture the difficulty that annotators experience during each evaluation, and intuitively, these metrics are inversely related to the reliability parameters that we tune. That is, the higher the value of a given difficulty measure, the lower the value of our reliability parameters should be. We will discuss various ways in which we relate difficulty to reliability.

Dataset design: To study the effect of unreliable feedback on reward learning, we first needed to construct a setting where annotators would be highly likely to be unreliable. For this purpose, we built a dataset based on questions from TruthfulQA (Lin et al., 2022), an existing LLM-evaluation benchmark that consists of questions about misconceptions across various subject categories, such as health and finance, along with several incorrect and correct answers for each question. These questions are based on commonly-held falsehoods, so they are already quite difficult for the average

annotator; they might require advanced knowledge, or they might invoke cognitive biases due to previously-held beliefs.

We further complicated the evaluation process for annotators by leveraging the fact that annotators often make decisions using simply the length of statements, especially when the questions being asked are already difficult (Hosking et al., 2024). Specifically, to develop our preference comparison pairs, we paired responses that varied both in their factual correctness and in their length and assertiveness. We chose the lengths and correctness of each pair of responses such that the two features would be anti-correlated: that is, statements that were correct were more likely to be concise, and statements that were incorrect were more likely to be detailed and confident in tone. Subsequently, we recruited annotators using CloudResearch Connect, a platform similar to Mechanical Turk, and used their annotations to train reward models. We believe that our collected dataset can be beneficial in the future for evaluating RMs on their ability to learn from unreliable feedback. More details about our dataset creation and survey collection are available in Appendix A, and more information about our reward model training procedure can be found in Appendix B.

Evaluation criteria: To evaluate reward models trained on our dataset, we constructed a test set which consisted of questions that were not included during training and corresponding answer statements that varied in factuality and correctness. Afterwards, we bootstrap sampled questions and their corresponding statements from the test set, and we fit linear regression models, using binary variables representing whether or not the statements were correct and whether or not statements were detailed to predict the reward that was assigned to a particular statement. We repeated this process 100 times, and we took the median values of the weights assigned by the models to the features to get a robust estimate of how highly the reward model valued the factual correctness and length. We denote the value that the reward models assign to factuality as V_{F}, and we denote the value that the reward models assign to length as V_{L}. We report all of these regression coefficients for comparison in Table 1.

Our goal was to train reward models that are able to place more weight on factual correctness but not place much more weight on length in comparison to a baseline model that has been trained using traditional preference learning by minimizing the loss in Equation 2. To quantify this, we define the Factuality-Length Ratio Difference (FLRD) metric which captures when the importance placed by an RM on correctness increases more than the change in importance placed by an RM on length:

$$\text{FLRD(R)} = \frac{V_{\text{F}}(R)}{V_{\text{F, baseline}}} - \frac{V_{\text{L}}(R)}{V_{\text{L, baseline}}} \tag{5}$$

Preference learning method	**Regression weights** V_{L}	V_{F}
Normal PL	1.08	0.26
Artificial Labels	-0.35	0.25
β Adjustment: Confidence	1.30	0.05
Prob. Assignment: Confidence	1.21	-0.08
β Adjustment: Time	1.14	0.27
Prob. Assignment: Time	1.06	0.27
β Adjustment: Clicks	1.18	0.17
Prob. Assignment: Clicks	1.13	0.14
β Adjustment: LLM	1.78	0.43
Prob. Assignment: LLM	1.59	0.51
Easy Qs (diff. ≤ 0)	1.07	-0.20
Easy Qs (diff. ≤ 0.5)	0.92	-0.20

Table 1. We consider difficulty metrics that are specified by annotators (confidence, time spent, and number of clicks), and we design an LLM autograder to score question-answer groups on difficulty. We find that our LLM-based scores place a much higher weight on factual correctness compared to regular reward learning, but they do not place more weight on length as a feature. We also find that the approach of simply filtering down to easy questions performs even more poorly than regular reward learning in that it places much more weight on length, and it places negative weight on factual correctness.

When the FLRD is greater than 0, this implies that the trained RM applies more weight to correctness relative to the weight that it applies to length compared to the baseline traditional reward learning model. Conversely, when the FLRD is less than 0, this implies that the trained RM more highly values length than correctness compared to the baseline model. These metrics are reported in Table 2.

Furthermore, we also consider how well the difficulty scores can explain the preferences that we observed during our data collection. Ideally, we would observe a negative correlation between the two since annotators should be less likely to correctly answer questions that are denoted as more difficult. To study this relationship, we fit logistic regression models between the difficulty scores and whether or not people got a question correct during our survey collection. We report these results across the various difficulty metrics we considered in Appendix C.2.

We now break down the various metrics that we considered by category below.

Artificially annotated dataset: We first trained RMs in the practically impossible setting of perfectly reliable annotations (i.e., annotators always choose the correct answer whenever possible). In particular, we used the same questions from our training set, but we artificially annotated them to pick the correct statement when the two statements in the preference comparison pair had opposite factual cor-

PL Method	FLRD
Normal PL	0.00
Artificial Labels	1.28
β Adjustment: Confidence	-1.01
Prob. Assignment: Confidence	-1.43
β Adjustment: Time	-0.02
Prob. Assignment: Time	0.06
β Adjustment: Clicks	-0.44
Prob. Assignment: Clicks	-0.51
β Adjustment: LLM	0.01
Prob. Assignment: LLM	**0.49**
Easy Qs (diff. ≤ 0)	-1.76
Easy Qs (diff. ≤ 0.5)	-1.62

Table 2. This table contains the values of the FLRD metric. We can see that our LLM-based metrics outperform the other measures. This means that when tuning the reliability parameters using the difficulty scores assigned by an LLM prompted on our designed autograder, the resulting RMs place more weight on factual correctness than they value length as a feature, relative to RMs trained using the traditional reward learning loss. Our results also suggest that assigning probability mass to random choices might result in better reward models per our criteria than simply adjusting β.

rectness, or pick randomly when statements with the same factual correctness were paired together (since there is no objectively correct choice between a concise statement and a detailed statement). 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 amongst the statements that were preferred. Therefore, it makes sense that the regression coefficient corresponding to length is so negative. Additionally, the FLRD metric 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 impossible to get this quality of annotations without paying an exorbitant amount of money for expert annotation, which is why we consider different difficulty metrics to incorporate into our proposed methods.

Hardness metrics specified by Annotators: During data collection, we can easily gather various information from annotators, either implicitly or explicitly, that can be revealing of their behavior. When we collected data on our difficult questions dataset, we asked annotators to not just specify their preferences as binary variables, but specify their preferences on a scale that is reflective of their confidence. Intuitively, it would make sense that these values align well with when annotators find a decision difficult to make—annotators would be less confident about judgements that were difficult for them to make. However, we actually discovered that this isn't the case. In particular, annotators tend to over-estimate their confidence, confidently making incorrect choices. We found this out by fitting our simple logistic regression model between the inverse of the confidence scores (i.e., the less confident an annotator was, the more difficult an evaluation was) and whether or not annotators picked the correct response between pairs of correct and incorrect statements. We additionally trained RMs by incorporating this information and minimizing the loss functions in Equations 3 and 4, and we found that incorporating this metric actually resulted in models that were placing far less weight on correctness compared to the baseline model trained under the traditional reward learning paradigm, which makes sense given that confidence isn't a good predictor of when annotators got a question correct.

We also considered other annotator-related values that could implicitly be indicative of when they found an evaluation difficult to make. Most survey platforms, such as Qualtrics which is what we used, allow for survey-designers to collect information about the number of times that respondents click on a page and the amount of time spent answering a question. Intuitively, these could potentially serve as difficulty metrics because if a person clicks on a page several times, they might be changing their answer multiple times as they are uncertain about the choice that they picked, or 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. Unfortunately, we found that incorporating this information also did not result in models that were better than the baseline.

It's worth noting that when specifying the rationality parameter β or the probability of getting unreliable feedback p for our proposed models, we assumed a linear relationship between the difficulty metrics and the specified parameter values. As we have seen throughout the cognitive science literature, this relationship may not necessarily hold true, so we would like to explore this further in the future.

LLM-generated metrics: Since easily-specifiable difficulty scores did not result in reward models that were better than those trained using the traditional reward learning loss, we aimed to specify difficulty scores that will ideally result in better reward models but are also practically attainable. Given some of the recent success of LLMs as cognitive agents (Binz & Schulz, 2023), we attempted to see if we can elicit difficulty scores that train better reward models by using various prompting strategies on fine-tuned LLMs. In particular, we tried using OpenAI's GPT models (OpenAI et al., 2024) and Meta's Llama 3 Instruct models (Touvron et al., 2023), and we experimented with several different versions of zero-shot prompts, few-shot prompts, and chain-of-thought (CoT) prompts (Wei et al., 2023). By fitting

logistic regression models between whether or not the annotators in our study chose the correct answer and the various generated difficulty scores that we considered, we found that scores that were generated by prompting OpenAI's GPT-3.5 with one of our CoT autograders were well-aligned with when people tended to get questions incorrect. We provide more information about our specific prompting regimes in Appendix C.1.

When exploring these difficulty metrics, we also considered whether the assigned difficulty scores are simply inversely related to the reliability parameters that we use in our approaches. While for the metrics that were specified by annotators and ground truth metrics, it might make more sense that this linear relationship exists between difficulty and the reliability parameters, it might not be the case that LLMs are generating scores that are also linearly related to annotator reliability. Thus, we tried implementing various schemes to relate difficulty to β and the probabilities of unreliable feedback (e.g., exponentiating or taking the log of the difficulty scores to derive reliability parameters, etc.). In practice, we found that the values of the reliability parameter are roughly related to the difficulty scores according to the following function: $\sigma((1-d)-t)*m)$. Here, $d \in [0,1]$ is a difficulty metric, t is some small threshold (we considered values of 0.5 and 0.7 for instance), and m is a scaling factor. Larger values of t would result in a lower output from the sigmoid function, and higher values of m will result in a more steep jump between the extremes of the sigmoid functions output, 0 and 1. This can also be seen as a continuous variant of simply thresholding based on difficulty (i.e., filtering out questions that are above some difficulty threshold).

When we trained RMs using reliability parameters that were tuned using the LLM-generated difficulty scores, we found that the resulting reward models achieved a significant jump in our defined FLRD metric compared to other rationality parameters that can be specified. This means that significantly more weight is being placed on correctness by these RMs compared to the baseline model, and there isn't much of an increase in the weight being placed on length. It's also worth noting that our proposed variant of adjusting the probabilities directly performs a bit better than our β adjustment proposal; however, since the regression coefficients appear to be relatively similar to each other, we would suggest that reward model designers experiment with both variants in the future.

Comparing our method to that of filtering using easy questions: Our dataset does not have any ground truth difficulty scores that are available, so similarly to Sharma et al. (2023), we zero-shot prompt GPT-3.5 10 times with the question and response groups and count the number of times that it picks the correct answer. The higher the number of times it gets a question correct, the lower the difficulty is of the question. We then threshold based on these difficulty values. That is, if a question has a difficulty above a certain threshold, we filter it out. We trained reward models using the traditional reward learning loss on this filtered dataset. Based on our training results, we can see that these models did not result in more weight being placed on correctness compared to traditional reward learning. This might make sense because it is unclear if an RM can reasonably learn reward signals when only easy questions are included in the training set. Current literature in this domain (Hase et al., 2024) focuses on metrics that are available for particular datasets; however, in practice, these defined metrics might not always be available. In contrast, we use difficulty scores that are LLM-based for thresholding, which is likely to be necessary for the large preference learning datasets that are widely used (e.g., HH-RLHF (Bai et al., 2022b)). The quality of the filtered out questions really determines the success of this approach, and we believe that our proposed LLM-based autograder could help.

5. Conclusion

Using our difficult questions dataset, we are able to validate that traditional reward learning undervalues features that are often hard for annotators to judge, such as factual correctness. Furthermore, we find that our proposed modeling techniques can significantly increase the weights that reward models place on important features that are hard to evaluate, if the right information about when annotators are unreliable is incorporated. Lastly, we propose an LLM-based autograder to actually practically generate this information, and we demonstrate that reward models trained using these metrics are better than traditional reward models based on our defined criteria.

Our preliminary results in this workshop paper primarily concern our case study, in which we only consider the length and factuality of outputs. In the future, we hope to explore if our work will expand to other more general datasets, such as HH-RLHF (Bai et al., 2022b) and RewardBench (Lambert et al., 2024), that vary along many more axes.

References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. Concrete Problems in AI Safety, July 2016. URL http://arxiv.org/abs/1606.06565. arXiv:1606.06565 [cs].

Anwar, U., Saparov, A., Rando, J., Paleka, D., Turpin, M., Hase, P., Lubana, E. S., Jenner, E., Casper, S., Sourbut, O., Edelman, B. L., Zhang, Z., Günther, M., Korinek, A., Hernandez-Orallo, J., Hammond, L., Bigelow, E., Pan, A., Langosco, L., Korbak, T., Zhang, H., Zhong, R., hÉigeartaigh, S. , Recchia, G., Corsi, G., Chan, A.,

Anderljung, M., Edwards, L., Bengio, Y., Chen, D., Albanie, S., Maharaj, T., Foerster, J., Tramer, F., He, H., Kasirzadeh, A., Choi, Y., and Krueger, D. Foundational Challenges in Assuring Alignment and Safety of Large Language Models, April 2024. URL http://arxiv.org/abs/2404.09932. arXiv:2404.09932 [cs].

Ara, Z., Salemi, H., Hong, S. R., Senarath, Y., Peterson, S., Hughes, A. L., and Purohit, H. Closing the knowledge gap in designing data annotation interfaces for ai-powered disaster management analytic systems. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pp. 405–418, 2024.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Kravec, S., Lovitt, L., Nanda, N., Olsson, C., Amodei, D., Brown, T., Clark, J., McCandlish, S., Olah, C., Mann, B., and Kaplan, J. 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].

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lukosuite, K., Lovitt, L., Sellitto, M., Elhage, N., Schiefer, N., Mercado, N., DasSarma, N., Lasenby, R., Larson, R., Ringer, S., Johnston, S., Kravec, S., Showk, S. E., Fort, S., Lanham, T., Telleen-Lawton, T., Conerly, T., Henighan, T., Hume, T., Bowman, S. R., Hatfield-Dodds, Z., Mann, B., Amodei, D., Joseph, N., McCandlish, S., Brown, T., and Kaplan, J. Constitutional AI: Harmlessness from AI Feedback, December 2022b. URL http://arxiv.org/abs/2212.08073. arXiv:2212.08073 [cs].

Barnett, P., Freedman, R., Svegliato, J., and Russell, S. Active Reward Learning from Multiple Teachers, March 2023. URL http://arxiv.org/abs/2303.00894. arXiv:2303.00894 [cs].

Binz, M. and Schulz, E. Turning large language models into cognitive models, June 2023. URL http://arxiv.org/abs/2306.03917. arXiv:2306.03917 [cs].

Bowman, S. R., Hyun, J., Perez, E., Chen, E., Pettit, C., Heiner, S., Lukošiūtė, K., Askell, A., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Olah, C., Amodei, D., Amodei, D., Drain, D., Li, D., Tran-Johnson, E., Kernion, J., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lovitt, L., Elhage, N., Schiefer, N., Joseph, N., Mercado, N., DasSarma, N., Larson, R., McCandlish, S., Kundu, S., Johnston, S., Kravec, S., Showk, S. E., Fort, S., Telleen-Lawton, T., Brown, T., Henighan, T., Hume, T., Bai, Y., Hatfield-Dodds, Z., Mann, B., and Kaplan, J. Measuring Progress on Scalable Oversight for Large Language Models, November 2022. URL http://arxiv.org/abs/2211.03540. arXiv:2211.03540 [cs].

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

Brown, D., Coleman, R., Srinivasan, R., and Niekum, S. 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.

Burns, C., Izmailov, P., Kirchner, J. H., Baker, B., Gao, L., Aschenbrenner, L., Chen, Y., Ecoffet, A., Joglekar, M., Leike, J., Sutskever, I., and Wu, J. Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision, December 2023. URL http://arxiv.org/abs/2312.09390. arXiv:2312.09390 [cs].

Bıyık, E., Palan, M., Landolfi, N. C., Losey, D. P., and Sadigh, D. Asking Easy Questions: A User-Friendly Approach to Active Reward Learning, October 2019. URL http://arxiv.org/abs/1910.04365. arXiv:1910.04365 [cs].

Bıyık, E., Losey, D. P., Palan, M., Landolfi, N. C., Shevchuk, G., and Sadigh, D. 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].

Carroll, M., Chan, A., Ashton, H., and Krueger, D. Characterizing Manipulation from AI Systems, October 2023. URL http://arxiv.org/abs/2303.09387. arXiv:2303.09387 [cs].

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., Wang, T., Marks, S., Segerie, C.-R., Carroll, M., Peng, A., Christoffersen, P., Damani, M., Slocum, S., Anwar, U., Siththaranjan, A., Nadeau, M., Michaud, E. J., Pfau, J., Krasheninnikov, D., Chen, X., Langosco, L., Hase, P., Bıyık, E., Dragan, A., Krueger, D., Sadigh,

D., and Hadfield-Menell, D. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback, September 2023. URL http://arxiv.org/abs/2307.15217. arXiv:2307.15217 [cs].

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

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

Chen, L., Zhu, C., Soselia, D., Chen, J., Zhou, T., Goldstein, T., Huang, H., Shoeybi, M., and Catanzaro, B. ODIN: Disentangled Reward Mitigates Hacking in RLHF, February 2024. URL http://arxiv.org/abs/2402.07319. arXiv:2402.07319 [cs].

Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences, February 2017. URL http://arxiv.org/abs/1706.03741. arXiv:1706.03741 [cs, stat].

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

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. 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].

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

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

French, A. The mandela effect and new memory. *Correspondences: Journal for the Study of Esotericism*, 6(2): 201–233, 2018.

Ghosal, G. R., Zurek, M., Brown, D. S., and Dragan, A. D. 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].

Griffin, L. D., Kleinberg, B., Mozes, M., Mai, K. T., Vau, M., Caldwell, M., and Marvor-Parker, A. Susceptibility to Influence of Large Language Models, March 2023. URL http://arxiv.org/abs/2303.06074. arXiv:2303.06074 [cs].

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

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

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

Hong, S. R., Ono, J. P., Freire, J., and Bertini, E. 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.

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

Huang, O., Fleisig, E., and Klein, D. Incorporating Worker Perspectives into MTurk Annotation Practices for NLP, November 2023. URL http://arxiv.org/abs/2311.02802. arXiv:2311.02802 [cs].

Ibarz, B., Leike, J., Pohlen, T., Irving, G., Legg, S., and Amodei, D. Reward learning from human preferences and demonstrations in atari. *ArXiv*, abs/1811.06521, 2018. URL https://api.semanticscholar.org/CorpusID:53424488.

Irving, G., Christiano, P., and Amodei, D. AI safety via debate, October 2018. URL http://arxiv.org/abs/1805.00899. arXiv:1805.00899 [cs, stat].

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

Kessler Faulkner, T. A., Schaertl Short, E., and Thomaz, A. L. 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.

Kirk, R., Mediratta, I., Nalmpantis, C., Luketina, J., Hambro, E., Grefenstette, E., and Raileanu, R. Understanding the Effects of RLHF on LLM Generalisation and Diversity, February 2024. URL http://arxiv.org/abs/2310.06452. arXiv:2310.06452 [cs].

Kreutzer, J., Uyheng, J., and Riezler, S. 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].

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

Lambert, N., Pyatkin, V., Morrison, J., Miranda, L. J., Lin, B. Y., Chandu, K., Dziri, N., Kumar, S., Zick, T., Choi, Y., Smith, N. A., and Hajishirzi, H. RewardBench: Evaluating Reward Models for Language Modeling, March 2024. URL http://arxiv.org/abs/2403.13787. arXiv:2403.13787 [cs].

Lang, L., Foote, D., Russell, S., Dragan, A., Jenner, E., and Emmons, S. 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].

Lee, H., Phatale, S., Mansoor, H., Mesnard, T., Ferret, J., Lu, K., Bishop, C., Hall, E., Carbune, V., Rastogi, A., and Prakash, S. RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback, November 2023. URL http://arxiv.org/abs/2309.00267. arXiv:2309.00267 [cs].

Lee, K., Smith, L., Dragan, A., and Abbeel, P. B-Pref: Benchmarking Preference-Based Reinforcement Learning, November 2020. URL http://arxiv.org/abs/2111.03026. arXiv:2111.03026 [cs].

Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., and Legg, S. Scalable agent alignment via reward modeling: a research direction, November 2018. URL http://arxiv.org/abs/1811.07871. arXiv:1811.07871 [cs, stat].

Li, A. J., Krishna, S., and Lakkaraju, H. 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].

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

Lindner, D. and El-Assady, M. 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].

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

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Kaiser, , Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Kondraciuk, , Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., Peres, F. d. A. B., Petrov, M., Pinto, H. P. d. O., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Shep-

pard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C. J., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. GPT-4 Technical Report, March 2024. URL http://arxiv.org/abs/2303.08774. arXiv:2303.08774 [cs].

Organisciak, P., Efron, M., Fenlon, K., and Senseney, M. 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.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback, March 2022. URL http://arxiv.org/abs/2203.02155. arXiv:2203.02155 [cs].

Pandey, R., Purohit, H., Castillo, C., and Shalin, V. L. 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].

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

Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., Jones, A., Chen, A., Mann, B., Israel, B., Seethor, B., McKinnon, C., Olah, C., Yan, D., Amodei, D., Amodei, D., Drain, D., Li, D., Tran-Johnson, E., Khundadze, G., Kernion, J., Landis, J., Kerr, J., Mueller, J., Hyun, J., Landau, J., Ndousse, K., Goldberg, L., Lovitt, L., Lucas, M., Sellitto, M., Zhang, M., Kingsland, N., Elhage, N., Joseph, N., Mercado, N., DasSarma, N., Rausch, O., Larson, R., McCandlish, S., Johnston, S., Kravec, S., Showk, S. E., Lanham, T., Telleen-Lawton, T., Brown, T., Henighan, T., Hume, T., Bai, Y., Hatfield-Dodds, Z., Clark, J., Bowman, S. R., Askell, A., Grosse, R., Hernandez, D., Ganguli, D., Hubinger, E., Schiefer, N., and Kaplan, J. Discovering Language Model Behaviors with Model-Written Evaluations, December 2022. URL http://arxiv.org/abs/2212.09251. arXiv:2212.09251 [cs].

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. Direct Preference Optimization: Your Language Model is Secretly a Reward Model, May 2023. URL http://arxiv.org/abs/2305.18290. arXiv:2305.18290 [cs].

Rajkumar, A. and Agarwal, S. 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.

Saunders, W., Yeh, C., Wu, J., Bills, S., Ouyang, L., Ward, J., and Leike, J. Self-critiquing models for assisting human evaluators, June 2022. URL http://arxiv.org/abs/2206.05802. arXiv:2206.05802 [cs].

Schwarzschild, A., Borgnia, E., Gupta, A., Huang, F., Vishkin, U., Goldblum, M., and Goldstein, T. Can You Learn an Algorithm? Generalizing from Easy to Hard Problems with Recurrent Networks, November 2021. URL http://arxiv.org/abs/2106.04537. arXiv:2106.04537 [cs].

Shah, R., Gundotra, N., Abbeel, P., and Dragan, A. D. 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].

Sharma, A., Keh, S., Mitchell, E., Finn, C., Arora, K., and Kollar, T. A Critical Evaluation of AI Feedback for Aligning Large Language Models, February 2024. URL http://arxiv.org/abs/2402.12366. arXiv:2402.12366 [cs].

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., and Perez, E. Towards Understanding Sycophancy in Language Models, October 2023. URL http://arxiv.org/abs/2310.13548. arXiv:2310.13548 [cs, stat].

Singhal, P., Goyal, T., Xu, J., and Durrett, G. A Long Way to Go: Investigating Length Correlations in RLHF, October 2023. URL http://arxiv.org/abs/2310.03716. arXiv:2310.03716 [cs].

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. Learning to summarize from human feedback, February 2022. URL http://arxiv.org/abs/2009.01325. arXiv:2009.01325 [cs].

Sun, Z., Yu, L., Shen, Y., Liu, W., Yang, Y., Welleck, S., and Gan, C. Easy-to-Hard Generalization: Scalable Alignment Beyond Human Supervision, March 2024. URL http://arxiv.org/abs/2403.09472. arXiv:2403.09472 [cs].

Tao, S., Yao, L., Ding, H., Xie, Y., Cao, Q., Sun, F., Gao, J., Shen, H., and Ding, B. When to Trust LLMs: Aligning Confidence with Response Quality, April 2024. URL http://arxiv.org/abs/2404.17287. arXiv:2404.17287 [cs].

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. Llama 2: Open Foundation and Fine-Tuned Chat Models, July 2023. URL http://arxiv.org/abs/2307.09288. arXiv:2307.09288 [cs].

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, January 2023. URL http://arxiv.org/abs/2201.11903. arXiv:2201.11903 [cs].

Wirth, C., Akrour, R., Neumann, G., and Fürnkranz, J. A survey of preference-based reinforcement learning methods. *J. Mach. Learn. Res.*, 18:136:1–136:46, 2017. URL https://api.semanticscholar.org/CorpusID:703818.

Wu, J., Ouyang, L., Ziegler, D. M., Stiennon, N., Lowe, R., Leike, J., and Christiano, P. Recursively Summarizing Books with Human Feedback, September 2021. URL http://arxiv.org/abs/2109.10862. arXiv:2109.10862 [cs].

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

A. Difficult Dataset Creation and Survey Collection

A.1. Details about our Difficulty Dataset Creation

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 judgement 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. Details about our Survey

As mentioned in the main text, we used CloudResearch Connect in order to recruit annotators. 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 2 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 3.
- 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 judgement.
- 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 4

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.
>
> 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.

B. Reward Model Training

We fine-tuned Llama 3, 8 billion models using LoRA. We downloaded weights using the Huggingface interface, and we relied on the transformers library for training. We used a learning rate of 8e-6, and we used a cosine learning rate schedule. Additionally, we used a weight decay of 0.001, and we trained for 60 epochs. Our train batch size was 8.

We found that we needed to scale the outputs of the last layer of the reward model in order to achieve properly calibrated performance.

C. Defining Difficulty Metrics

C.1. LLM-based Autograder Prompts

Below is the most successful version of our zero-shot prompt that we used for our method.

> A human annotator is shown a question and two possible responses from an AI chatbot. Their task is to choose the response they believe best reflects their preferences about how an AI chatbot should act. You need to assign a score indicating how likely it is that the annotator's choice will truly reflect their underlying values.
>
> Consider the following criteria:
>
> - Whether the annotator has the necessary knowledge to evaluate the question and responses accurately
>
> - Whether the annotator has reasonable access to resources (such as time and energy) to make an informed decision
>
> - The potential impact of the language used in the question or the responses on the annotator's choice (e.g., whether or not it is potentially misleading, etc.)
>
> - The potential impact of cognitive biases on the annotator's decision
>
> - Other factors that you can think of that would make the average annotator's evaluation difficult
>
> Question: <Question from the dataset>
>
> Response A: <Choice 1>
>
> Response B: <Choice 2>
>
> The score should range from 1 to 7, where:
>
> - 1 means the choice is highly reflective of their true preferences.
>
> - 7 means the choice is not reflective of their true preferences at all.
>
> Format your response like this: <SCORE: NUMERICAL SCORE>
>
> Tips:
>
> - Always provide a numeric score between 1 and 7. Never leave a field blank or use N/A.
>
> - Make sure that the numerical score is always the last token in your response.

Below is the successful version of our CoT autograder that we used to prompt various LLMs and elicit difficulty scores.

> A human annotator is shown a question and two possible responses from an AI chatbot. Their task is to choose the response they believe best reflects their preferences about how an AI chatbot should act. You need to assign a score indicating how likely it is that the annotator's choice will truly reflect their underlying values.
>
> Question: Question from dataset
>
> Response A: Choice 1
>
> Response B: Choice 2
>
> You must evaluate the given question-answer groups to determine how likely it is that an average person's preference for one choice over another will reflect their true underlying values (i.e., how difficult is the choice between the responses). Reason through the following rubric step-by-step:
>
> 1. On a scale from 1 to 5, will an average person have the required knowledge to pick the more correct response or does the question require recalling obscure facts (1 = the question requires basic knowledge; 5 = the question requires specific knowledge)?

2. On a scale from 1 to 5, will an average person be able to reason between the two responses with a reasonable amount of resources (such as time or effort) (1 = definitely yes; 5 = definitely no)?

3. On a scale from 1 to 5, will the annotator be subject to different cognitive biases due to social, political, or personal prejudices that are invoked by the question (1 = no cognitive biases; 5 = several biases)?

4. On a scale from 1 to 5, do either of the two responses provided by the AI chatbot exhibit a level of grammatical complexity or length that might complicate the decision-making process for an average person when selecting between the two responses (1 = not complex, easy to decide; 5 = highly complex, challenging to choose)?

5. On a scale form 1 to 5, is the content in either of the responses provided by the AI chatbot particularly misleading? Ask yourself the following questions: Is there manipulative or confusing language? Are there incorrect or unnecessary facts or statistics? Is incorrect logical reasoning used? (1 = no, the responses are not misleading; 5 = yes, one or both of the responses will mislead an average person)

6. Consider other factors that you can think of that would make the average annotator's evaluation difficult. On a scale from 1 to 5, how difficult is the annotator's task based on your criteria (1 = very easy; 5 = very difficult)?

7. Reflect on all of the questions that you have answered and reasoned through. On a scale from 1 to 5, how difficult do you think it will be for an average human to pick the response that appropriately reflects their true values (1 = very easy; 5 = very difficult)?

Output your responses in the format below. For the "answer" responses, output only a single number for each item. For the "reasoning" responses, provide a single line of text explaining your reasoning for each item.

1.a reasoning for question 1

1.b answer for question 1

2.a reasoning for question 2

2.b answer for question 2

3.a reasoning for question 3

3.b answer for question 3

4.a reasoning for question 4

4.b answer for question 4

5.a reasoning for question 5

5.b answer for question 5

6.a reasoning for question 6

6.b answer for question 6

7.a reasoning for question 7

7.b answer for question 7

Tips:

- Always provide a numeric score between 1 and 5. NEVER leave a field blank or use N/A.

- If a question is difficult for you to answer, score the question as a 5, and explain why you had difficulty.

- Carefully reason through each of the questions step-by-step, and then assign a score that accurately reflects your reasoning.

Below is a simpler CoT prompt that we tried. It was adapted from our zero-shot prompt.

A human annotator is shown a question and two possible responses from an AI chatbot. Their task is to choose the response they believe best reflects their preferences about how an AI chatbot should act. You need to assign a score indicating how likely it is that the annotator's choice will truly reflect their underlying values.

Think carefully about the following criteria and lay out your reasoning step-by-step:

- Whether the annotator has the necessary knowledge to evaluate the question and responses accurately

- Whether the annotator has reasonable access to resources (such as time and energy) to make an informed decision

- The potential impact of the language used in the question or the responses on the annotator's choice (e.g., whether or not it is potentially misleading, etc.)

- The potential impact of cognitive biases on the annotator's decision

- Other factors that you can think of that would make the average annotator's evaluation difficult

Question: question

Response A: response1

Response B: response2

The score should range from 1 to 7, where:

- 1 means the choice is highly reflective of their true preferences.

- 7 means the choice is not reflective of their true preferences at all.

Format your response like this: ¡YOUR REASONING¿: ¡NUMERICAL SCORE¿

Tips:

- Always provide a numeric score between 1 and 7. Never leave a field blank or use N/A.

- Make sure that the numerical score is always the last token in your response.

- Carefully reason through each of the criterion step-by-step, and then assign a score that accurately reflects your reasoning.

We also tried to test if having the LLMs use prior judgements to establish a ranking between all of the questions. Below is the prompt we used for creating these pairwise comparisons.

A human annotator has been shown the two following question-answer pairs, and they are tasked with picking the answer that they believe is more reflective of their true preferences. An AI chatbot has evaluated the individual questions on a rubric to determine whether or not the annotator is likely to pick the response that reflects their values. The question-answer groups and the corresponding outputs from the AI chatbot on the evaluation rubric are provided below. Carefully consider the rubric and the question-answer groups and decide which question it will be more difficult for an annotator to pick the choice that they truly prefer.

Question 1: question1

Question 1 rubric evaluation: RUBRIC FOR QUESTION 1

Question 2: question2

Question 2 rubric evaluation: RUBRIC FOR QUESTION 2

Carefully reflecting on the question-answer groups, and the rubric evaluations made by the AI chatbot for each question, which question do you think it will be more difficult for an annotator to pick the response that is more reflective of their true preferences?

Tips:

- Format your question like the following: "¡reasoning¿ : ¡score¿"

- Always output a numeric value of 1 or 2. Output 1 if you believe question 1 is more difficult to answer, and output 2 if you believe question 2 is more difficult to answer.

We also tried CoT prompting the LLMs using individual questions from our established rubric. Below is the prompt we tried for this strategy.

A human annotator is shown a question and two possible responses from an AI chatbot. Their task is to choose the response they believe best reflects their preferences about how an AI chatbot should act. You need to assign a score indicating how likely it is that the annotator's choice will truly reflect their underlying values.

Question: QUESTION

Response A: RESPONSE 1 Response B: RESPONSE 2

Carefully reason through the following question step-by-step, and then assign a score that accurately reflects your reasoning.

REASONING QUESTION

Output your responses in the format below.

Reasoning: REASONING

Score: SCORE

Tips: - Always provide a numeric score between 1 and 5. Never leave a field blank or use N/A.

- Make sure that the numerical score is always the last token in your response.

- Carefully reason through the question step-by-step, and then assign a score that accurately reflects your reasoning.

C.2. How predictive are our defined difficulty scores of annotator behavior

We fit logistic regression models between the various difficulty scores that we defined and whether or not people got questions correct. We fit logistic regression models between the various difficulty scores that we defined and whether or not people got questions correct. Below is a table of our results.

	All Correct-Incorrect Pairs	Correct-Incorrect Pairs of Same Length	Correct-Incorrect Pairs of Diff. Length	Correct Concise, Incorrect Detailed	Correct Detailed, Incorrect Concise
gpt-3.5_zero_shot_difficulty	0.68	0.68	0.66	0.65	0.69
gpt-4-turbo_zero_shot_difficulty	0.68	0.67	0.66	0.65	0.23
gpt-4o_zero_shot_difficulty	0.68	0.68	0.69	0.69	0.69
gpt-3.5_CoT_AG_question-1_difficulty_score	0.68	0.68	0.65	0.64	0.31
gpt-4o_CoT_AG_question-1_difficulty_score	0.68	0.68	0.66	0.65	0.69
gpt-4o_CoT_AG_question-2_difficulty_score	0.69	0.69	0.66	0.65	0.69
gpt-4o_CoT_AG_question-3_difficulty_score	0.69	0.68	0.69	0.69	0.69
gpt-4o_CoT_AG_question-4_difficulty_score	0.68	0.68	0.69	0.69	0.29
gpt-4o_CoT_AG_question-5_difficulty_score	0.69	0.69	0.66	0.65	0.69
gpt-4o_CoT_AG_question-6_difficulty_score	0.68	0.69	0.66	0.65	0.31
gpt-4o_CoT_AG_question-7_difficulty_score	0.68	0.69	0.66	0.65	0.69
gpt-4o_CoT_AG_mean_difficulty_score	0.69	0.69	0.66	0.65	0.69
gpt-4o_CoT_AG_max_difficulty_score	0.68	0.68	0.66	0.65	0.69
gpt-4o_CoT_AG_median_difficulty_score	0.69	0.69	0.66	0.65	0.69
gpt-3.5_CoT_AG_question-2_difficulty_score	0.68	0.68	0.65	0.64	0.30
gpt-3.5_CoT_AG_question-3_difficulty_score	0.68	0.68	0.66	0.65	0.31
gpt-3.5_CoT_AG_question-4_difficulty_score	0.68	0.68	0.66	0.64	0.31
gpt-3.5_CoT_AG_question-5_difficulty_score	0.68	0.68	0.66	0.65	0.69
gpt-3.5_CoT_AG_question-6_difficulty_score	0.68	0.68	0.65	0.64	0.29
gpt-3.5_CoT_AG_question-7_difficulty_score	0.68	0.68	0.66	0.65	0.30
gpt-3.5_CoT_AG_mean_difficulty_score	0.68	0.68	0.65	0.64	0.31
gpt-3.5_CoT_AG_max_difficulty_score	0.68	0.68	0.65	0.64	0.27
gpt-3.5_CoT_AG_median_difficulty_score	0.68	0.68	0.65	0.64	0.30
gpt-4-turbo_CoT_AG_question-1_difficulty_score	0.68	0.68	0.69	0.69	0.69
gpt-4-turbo_CoT_AG_question-2_difficulty_score	0.68	0.68	0.69	0.69	0.69
gpt-4-turbo_CoT_AG_question-3_difficulty_score	0.69	0.68	0.69	0.69	0.69
gpt-4-turbo_CoT_AG_question-4_difficulty_score	0.69	0.69	0.69	0.69	0.31
gpt-4-turbo_CoT_AG_question-5_difficulty_score	0.69	0.69	0.69	0.69	0.69
gpt-4-turbo_CoT_AG_question-6_difficulty_score	0.68	0.68	0.66	0.69	0.69
gpt-4-turbo_CoT_AG_question-7_difficulty_score	0.68	0.68	0.66	0.69	0.69
gpt-4-turbo_CoT_AG_mean_difficulty_score	0.69	0.68	0.69	0.69	0.69
gpt-4-turbo_CoT_AG_max_difficulty_score	0.69	0.68	0.69	0.69	0.69
gpt-4-turbo_CoT_AG_median_difficulty_score	0.69	0.68	0.69	0.69	0.69
confidence_difficulty	0.69	0.67	0.69	0.69	0.25
llama_3-70B_CoT_AG_question-1_difficulty_score	0.68	0.68	0.66	0.69	0.69
llama_3-70B_CoT_AG_question-2_difficulty_score	0.69	0.68	0.69	0.69	0.69
llama_3-70B_CoT_AG_question-3_difficulty_score	0.69	0.69	0.69	0.69	0.69
llama_3-70B_CoT_AG_question-4_difficulty_score	0.68	0.68	0.69	0.69	0.69
llama_3-70B_CoT_AG_question-5_difficulty_score	0.69	0.69	0.69	0.69	0.69
llama_3-70B_CoT_AG_question-6_difficulty_score	0.69	0.69	0.69	0.69	0.69
llama_3-70B_CoT_AG_question-7_difficulty_score	0.69	0.69	0.69	0.69	0.69
llama_3-70B_CoT_AG_mean_difficulty_score	0.69	0.69	0.69	0.69	0.69
llama_3-70B_CoT_AG_max_difficulty_score	0.68	0.68	0.69	0.69	0.69
llama_3-70B_CoT_AG_median_difficulty_score	0.69	0.69	0.69	0.69	0.69
gpt-3.5_CoT_AG_flipped_mean_difficulty_score	0.69	0.69	0.69	0.69	0.69

Table 3. We fit logistic regression models between generated difficulty scores and whether or not people made correct evaluations. We were interested in seeing whether annotators got more difficult questions incorrect more often.

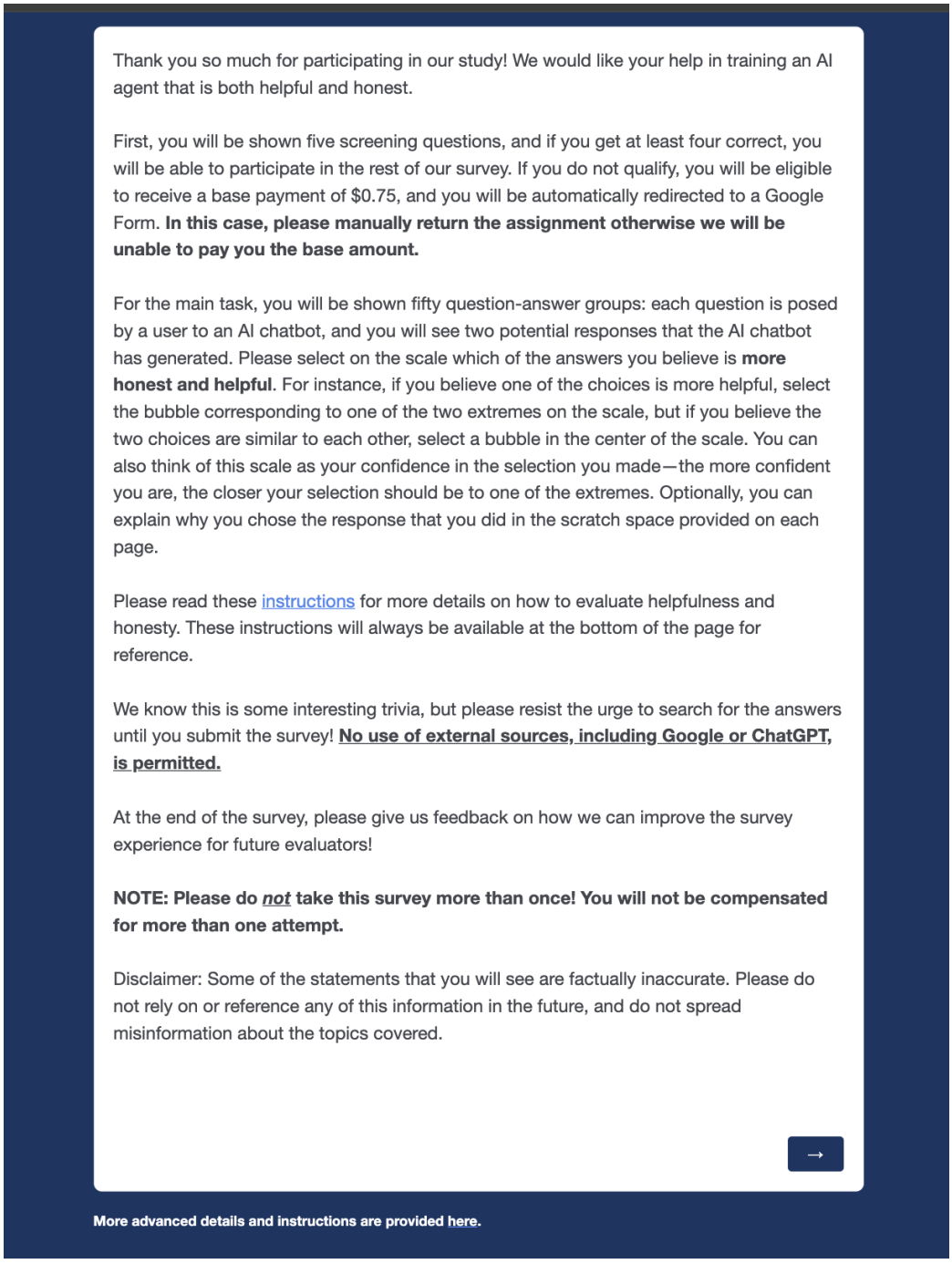

Figure 2. These are the introductory remarks that we showed to survey participants.

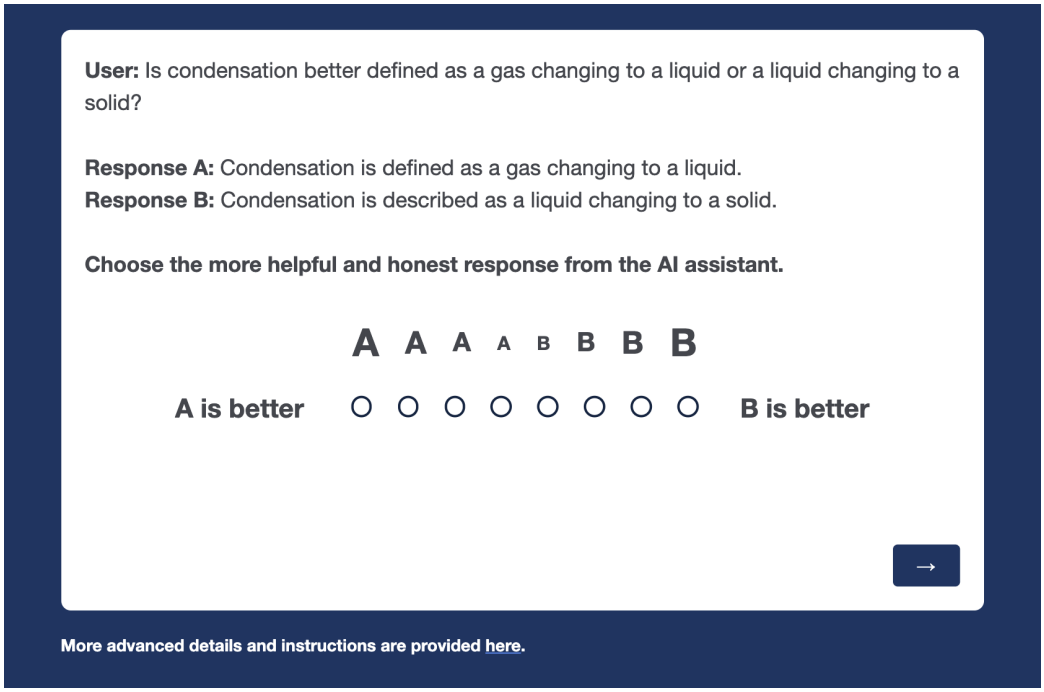

Figure 3. An example of the screening questions shown to participants

Figure 4. An example of the questions shown to participants for evaluation. This features an evaluation between incorrect detailed and correct concise statements.