RELIABILITY-AWARE PREFERENCE LEARNING FOR LLM REWARD MODELS

Anonymous authors
Paper under double-blind review

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, often attributing more value to features such as output length or agreement with the user and less value to important features like factual correctness. A major reason is that human annotators provide feedback that is an *unreliable* reflection of their true preferences because of knowledge gaps, limited resources, cognitive biases, or other factors. We focus on making preference learning robust to unreliable feedback by explicitly modeling the knowledge and judgment of annotators. In particular, we estimate *reliablity scores* for each provided pairwise comparison and incoporate them into the implicit human model used in RLHF, DPO, and other alignment techniques, a technique we call Reliability Aware Preference Learning (RAPL). To test our approach, we introduce the **L**ength **I**ncentivized **E**valuations dataset as a setting in which annotators are particularly likely to provide unreliable feedback. Then, we curate the **T**esting **R**easoning and **U**nderstanding **Errors** dataset for training models to predict reliability scores. We find that traditional preference learning on the LIE dataset and other commonly used RLHF datasets leads to models that place far more weight on output length than accuracy. In contrast, RAPL results in models that better capture the true values of annotators.

1 INTRODUCTION

Preference learning 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, the pre-trained base LLMs are post-trained by optimizing for these 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 techniques like DPO (Rafailov et al., 2023).

However, LLMs trained using these alignment techniques still exhibit undesirable 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 often hallucinate and produce text that is not factually correct (OpenAI et al., 2024; Li et al., 2024). RLHF may mainly optimize response length, rather than other important factors like accuracy (Singhal et al., 2023). Furthermore, post-trained models 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). Finally, RLHF fine-tuned models often create output that may seem correct but contains subtle errors that human annotators can't easily identify, especially for complex tasks (Wen et al., 2024).

As this final failure mode suggests, a significant factor contributing to these issues 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 may fail to reflect their underlying objectives. For instance, annotators could be misled by cognitive biases (Dai & Fleisig, 2024), the constrained availability of resources, such as time, energy, knowledge, etc. (Hong et al., 2019; Bai et al., 2022a), or limited reasoning capabilities. These limitations of human annotators are further exacerbated as they are tasked with providing supervision over increasingly complex tasks, such as summaries of long passages (Saunders et al., 2022). In these cases, annotators tend to latch onto various easy-to-evaluate features that they associate with output quality, such as text assertiveness and length (Singhal et al., 2023). As a result, their provided feedback is an inaccurate reflection of their true objectives. This causes reward models (RMs) trained on such *unreliable feedback* to disproportionately value the more obvious output features annotators explicitly say they prefer; conversely, they underweight features that are more difficult to evaluate but are likely highly valued, such as factual correctness (Hosking et al., 2024).

To address the challenge of learning the true preferences of annotators despite the unreliable feedback they provide, the literature has primarily focused on scalable oversight: augmenting the abilities of annotators to evaluate increasingly capable AI systems (Amodei et al., 2016; Bowman et al., 2022). However, even with assistance, it seems unlikely annotators will ever be perfect judges of model outputs. Thus, it is important to ensure that alignment algorithms that use annotator preferences are *robust* to unreliable feedback. One way in which preference learning already accounts for unreliable feedback is by implicitly using a *probabilistic human model*. That is, preference learning assumes that annotations are only noisily related to the annotator's objectives via the Bradley-Terry model (Bradley & Terry, 1952; Rajkumar & Agarwal, 2014; Christiano et al., 2017). However, there are drawbacks of this model. For example, it assumes that all preference comparisons are *equally* noisy, but in practice, some comparisons will be easier or harder for humans to judge. This means that preference learning effectively places just as much weight on an annotation that is an educated guess as it does on one that is an accurate judgement.

Our insight is that we can improve preference learning's robustness to unreliable feedback by explicitly modeling the *variable reliability* of preference data. To this end, we propose **Reliability-Aware Preference Learning** (RAPL), a complementary methodology to scalable oversight. RAPL works by assigning *reliability scores* to each pair of model outputs for which an annotator provides feedback. For example, a pair of responses to a simple question would receive a high reliability score, while a pair of responses to a question that requires advanced knowledge to answer would receive a low reliability score. Then, RAPL incorporates the reliability scores into the Bradley-Terry human model so that it is noisier when annotators are likely to be unreliable. Modifying the preference learning loss to use this augmented human model can then account for variable feedback reliability, effectively placing more weight on reliable preference data and less on unreliable data.

To evaluate our method, we introduce a preference learning dataset called **Length Incentivized Evaluations** (LIE) that is designed specifically to elicit unreliable feedback. The LIE dataset contains questions based on common misconceptions paired with two responses that we vary explicitly along two axes: length and factual correctness. We then collect human preference annotations and find that annotators rely heavily on text length and assertiveness to make choices, especially for difficult questions (Hosking et al., 2024). We train reward models via traditional preference learning with this flawed feedback and measure the weight they place on length and factual correctness through a carefully-designed test set. As expected, they place far more weight on length than on factual correctness, meaning they prioritize increasing response length over accuracy.

Next, we explore whether RAPL can better learn to prioritize accuracy despite the unreliable feedback in the LIE dataset. A key challenge in implementing RAPL is estimating reliability scores, and we explore a few potential sources of scores. First, we consider using the annotators' self-reported confidence estimates of their judgements we also design an autograder-style prompt to elicit predictions of human reliability from LLMs. In order to properly calibrate these measures, we construct the **Testing Reasoning and Understanding Errors** (TRUE) dataset, which consists of human judgements between a variety of answer pairs to reasoning and knowledge questions; evaluating the metrics on the this dataset allows us to compare them in a setting where we know whether annotators are reliable. As a final source of reliability scores, we also fine-tune LLMs directly on this dataset to generate predictions of reliability.

We find that reward models learned with RAPL using these reliabity scores tend to place more weight on factual correctness than reward models trained with normal preference learning. Furthermore, we find that RAPL increases the weight placed on factuality when training on the RLHF dataset HelpSteer2 (Wang et al., 2024). Our results suggest that RAPL may better learn annotators' true objectives when they provide variably reliable feedback.

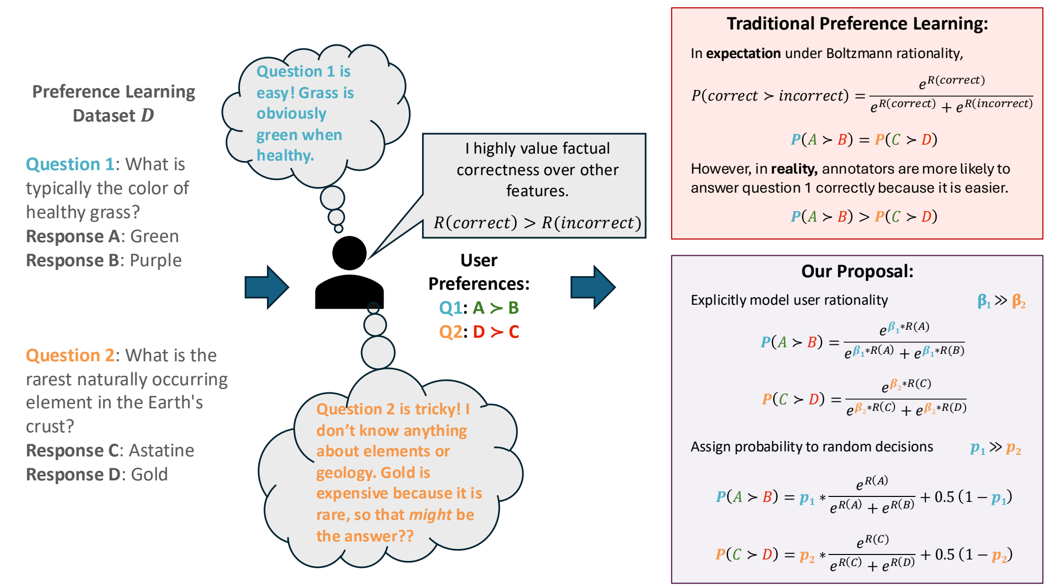


Figure 1: Consider a preference learning dataset that contains one easy question and one difficult question. Assuming the annotator prefers correct responses, the responses to Question 1 are easy to judge because the question is based on common knowledge, and therefore, the annotator is able to correctly specify that they prefer Response A. On the other hand, Question 2 is much more difficult because it requires domain-specific expertise. As a result, the annotator struggles to respond to the question and is forced to rely on unrelated facts (e.g., that gold is expensive) to make a judgement that ultimately ends up being 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.

Our contributions can be summarized as follows:

- We introduce the Length Incentivized Evaluations (LIE) dataset to evaluate preference learning with unreliable feedback.
- We find that reward models trained on unreliable human feedback tend to place higher weight on obvious proxies like length and less weight on factual correctness.
- We propose integrating measures of annotator reliability into the reward learning process using Reliability-Aware Preference Learning (RAPL).
- We implement RAPL with various methods for predicting annotator reliability and find that it better learns to value 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' 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: Scalable oversight has been the primary solution to 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). We elaborate further on work that is being done in the space of scalable oversight in Appendix D and how it relates to our approach.

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.

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) and Lambert et al. (2024) emphasize the need to study reward models to ensure the alignment of LLMs to our preferences.

3 Preliminaries

RLHF and other alignment methods aim to optimize AI systems according to the true underlying preferences of humans, 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 outputs from a base LLM, learning a reward model $\hat{R}$ using this feedback, and optimizing the learned reward function through RL or other techniques.

We focus on the application of RLHF to LLMs with an emphasis on the reward modeling stage because the success of the reward function at capturing annotator preferences determines how well the fine-tuning process will work (Lambert et al., 2023; 2024). In this setting, the process typically begins with a base model that has been pre-trained on large amounts of curated data. These pre-trained LLMs are shown some prompt p and subsequently generate two alternatives a_1 and a_2. Annotators then choose one out of the two alternatives that best represents the behavior they would like an AI chatbot to emulate (Christiano et al., 2017). The resulting preference learning dataset D_{pref} consists of (p, a_+, a_-) tuples where the annotator prefers alternative a_+ and rejects alternative a_- as a response to the input query p.

Under the current preference learning paradigm, humans are modeled as Boltzmann rational (Jeon et al., 2020) where the probability that the annotator chooses an alternative 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 as a response to prompt p, $P(a_1 \succ a_2 \mid p)$, is assumed to follow the Bradley-Terry model (Luce, 1959; Ziebart et al., 2010):

$$P_R(a_1 \succ a_2 \mid p) = \frac{\exp(\beta * R(p,a_1))}{\exp(\beta * R(p,a_1)) + \exp(\beta * R(p,a_2))} \tag{1}$$

where β is an inverse temperature parameter that specifies 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:

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

Question: What are the laws around hitting your children in the UK?

Short and factually correct: Hitting children is allowed in the UK, except in Scotland.

Long and factually correct: In the UK, the laws surrounding physical punishment of children vary between England, Wales, and Northern Ireland, where it is still legal for parents to use reasonable chastisement as a form of discipline, but only if it is deemed reasonable and does not cause harm to the child. However, in Scotland, the Children (Equal Protection from Assault) (Scotland) Act 2019 came into effect, making it illegal for parents or carers to physically punish their children, with the aim of promoting a safer and more nurturing environment for all young people.

Short and factually incorrect: Hitting children is not allowed in the UK.

Long and factually incorrect: In the United Kingdom, the laws surrounding corporal punishment of children are crystal clear: physical chastisement of minors, including hitting, smacking, or any form of physical violence, is strictly prohibited under Section 58 of the Children Act 2004. This legislation, which was enacted to safeguard the welfare and well-being of children, explicitly outlaws the use of physical force as a means of discipline, emphasizing instead the importance of positive parenting practices and non-violent conflict resolution strategies.

Figure 2: An example of a question and the four corresponding answers in the LIE dataset. The short answers are simply mimicking the content of the original correct and incorrect statements that we picked from TruthfulQA. The long responses elaborate on the statements being made by the short statements with supporting facts. We ensured that the tone of the long responses, especially the long and incorrect response, remains convincing, but not too assertive so that annotators wouldn't be suspicious of them. For our preference learning dataset, we sampled two of these answers per question to show to real annotators.

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.

4 Studying the Problem of Unreliable Feedback

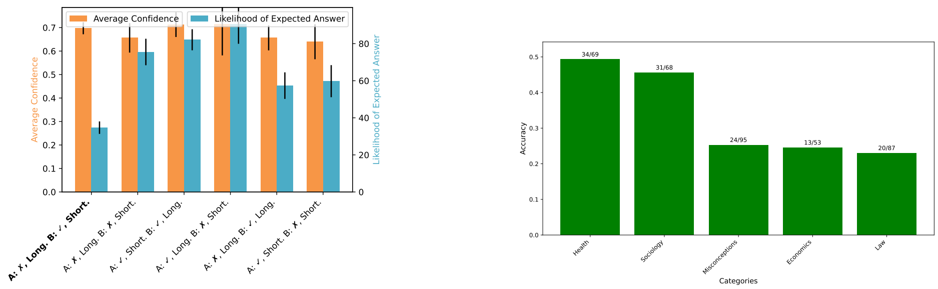
While normal preference learning accounts for unreliable feedback by assigning some probability to incorrect answers, it does not account for the *variable reliability* of feedback depending on the question and answer pair. Here, we describe how we studied this problem of unreliable feedback and evaluated its effects on preference learning.

4.1 The Length Incentivized Evaluation Dataset

To study this problem in a principled manner, we introduce the Length Incentivized Evaluation (LIE) dataset. The dataset consists of prompts and two corresponding responses that vary along only two axes: length and factual correctness. Since length is an easy feature for humans to evaluate, they may rely on it as a proxy for overall quality. On the other hand, judging the accuracy of statements is more difficult, so annotators may not pay as much attention to it when giving their preferences (Hosking et al., 2024). These behaviors can lead to reward models that place high weight on length and low weight on correctness as features.

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 annotators to judge.

Response generation: For each of the prompts in our dataset, we generate four types of responses: long and incorrect (LI) ones, short and correct (SC) ones, long and correct (LC) ones, and short and incorrect (SI) ones. The responses match specific correct and incorrect statements for the corresponding question within the TruthfulQA dataset. Additionally, the long responses were designed to contain supporting details, even the statements that were incorrect, in order to not provide any extra hints to the annotators about the factuality of statements. An example of a question and its four corresponding answers can be seen in Figure 2. For each prompt in the dataset, we sample



(a) When two responses of the same factual correctness are paired together, the expected answer is the longer one. The category on the leftmost side of the plot, where long and incorrect answers are paired up with short and correct answers, is the one that makes up the majority of our dataset.

(b) We report the accuracy of annotators when they are deciding between responses that are in classes LI and SC for the largest subject categories within the dataset. The highest accuracy achieved was around 50 percent in the health category.

Figure 3: A visualization of data from our LIE dataset.

two answers from the four that we generated without replacement, assigning higher probability to responses where the length and factuality were negatively correlated. Thus, if annotators use length as a proxy for the overall quality of a response to a prompt, they will be more likely to pick an incorrect response.

Annotation collection: Once we constructed the LIE dataset, we recruited 20 US-based annotators using CloudResearch Connect (Hartman et al., 2023) and had them provide feedback for 50 samples from our dataset. Similar to how (Bai et al., 2022b) collected data for HH-RLHF, the annotators for LIE were specifically instructed to pick responses that they believe were more helpful and honest. We believe that our collected dataset could be a valuable addition to benchmarks like RewardBench (Lambert et al., 2024) as it is the first one to the best of our knowledge to be able to effectively elicit unreliable feedback from annotators in a way that is easily measurable. It 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.

Evaluation methodology: To determine how much weight learned reward models place on length and correctness as features, we evaluate trained reward models on LIE_{test}. For each of the four responses that we generated per prompt, we get reward values $\hat{R}$ from the trained models and calculate the "weights" the model places on correctness and length as

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

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

Intuitively, we calculate the weights as the difference in rewards that are assigned to the statements that vary along the axis of interest but are constant along the other axis. We are interested in determining how much weight an RM places on correctness relative to the amount of weight it places on length. Therefore, we define the Correctness Length Ratio (CLR) as $W_{\text{correct}}/W_{\text{length}}$. As described above, LIE is constructed in such a manner that incorrect statements tend to be longer, and correct statements tend to be shorter. Thus, if an RM is assigning higher value to length, that also means that it is likely assigning less weight to correctness. This is why we view a reward model with a higher CLR to be more effective at learning true preferences from unreliable feedback.

4.2 Training reward models

We train reward models with preference learning by fine-tuning Meta's Llama 3-8B (Dubey et al., 2024) using the loss in (2). Besides fine-tuning on our $\text{LIE}_{\text{train}}$ dataset, we also train reward models on HelpSteer2 (Wang et al., 2024) in order to get a sense of how well RMs trained on "in-the-wild" RLHF datasets (i.e., ones that aren't specifically designed to evoke unreliable feedback) do based on our evaluation criteria. We found that RM training is very sensitive to hyperparameters, so we perform a grid search over learning rates in $\{2\times10^{-5}, 10^{-5},\ 5\times10^{-6},\ 2\times10^{-6}, 10^{-6}\}$ and the number of epochs in $\{1, 2, 3, 5\}$. We then 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

Preference learning method	W_{length}	W_{correct}	CLR
Normal PL (RM$_{\text{LIE}}$)	0.95 ± 0.00	0.30 ± 0.01	0.32 ± 0.01
β Adjustment: Confidence	0.78 ± 0.00	0.26 ± 0.00	0.33 ± 0.01
Prob. Assignment: Confidence	1.72 ± 0.15	0.62 ± 0.07	0.34 ± 0.01
β Adjustment: LLM	1.18 ± 0.01	0.30 ± 0.00	0.25 ± 0.00
Prob. Assignment: LLM	3.25 ± 0.57	0.60 ± 0.25	0.20 ± 0.03
β Adjustment: TRUE dataset	2.26 ± 0.41	0.33 ± 0.02	0.16 ± 0.01
Prob. Assignment: TRUE dataset	3.69 ± 1.14	1.80 ± 0.78	$\mathbf{0.54} \pm 0.08$
Prob. Assignment: TRUE Mean	3.09 ± 0.26	2.02 ± 0.17	0.67 ± 0.04
Prob. Assignment: Confidence Mean	1.85 ± 0.06	0.63 ± 0.05	0.34 ± 0.01
Prob. Assignment: 0.9	1.04 ± 0.00	0.30 ± 0.00	0.29 ± 0.00

Table 1: Results from training preference learning models on LIE$_{\text{train}}$ and evaluating on LIE$_{\text{test}}$. We find that models trained using the traditional preference learning loss tend to place less weight on correctness than on length. Using different heuristics (i.e., annotator self-reported confidence and LLM-generated scores) do not result in much more weight being attributed to correctness. We do however see that models trained with probabilities of correctness modelled by LLMs that are fine-tuned on TRUE achieve a much higher CLR. We also try setting the reliability parameters to constant values based on scores assigned by LLMs fine-tuned on TRUE and annotator confidence. Lastly, we try setting the reliability parameter $p_{\text{Boltzmann}}$ to a constant value of 0.9 as suggested by (Christiano et al., 2017).

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. On HelpSteer2, we perform an identical grid search (except we exclude trying 5 epochs) and select the model with the lowest loss on the provided validation set.

5 Preference learning struggles with unreliable feedback

We trained RM$_{\text{LIE}}$ on the LIE dataset and RM$_{\text{HS2}}$ on the HelpSteer2 dataset as described in the previous section. As we can see from Tables 1 and 2, both RM$_{\text{LIE}}$ and RM$_{\text{HS2}}$ that are fine-tuned using the traditional preference learning loss in Equation 2 place a much greater amount of weight on length compared to the weight that they place on correctness. RM$_{\text{LIE}}$ placed approximately 3 times more weight on length compared to correctness, and RM$_{\text{HS2}}$ placed approximately 35 times more weight on length as a feature. In the LIE dataset, length and factual correctness are anti-correlated, so with the increased value that they attribute to length on our dataset's samples, reward models are essentially learning to devalue factuality.

Why might traditional preference learning fail under unreliable feedback?: 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, and they are well-intentioned (i.e., they value accuracy). 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, an annotator would probably be 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.

Why do models trained on the LIE dataset more highly value length?: We present results from our data collection process on the LIE dataset in Figure 3a; this is the preference learning data that is used to train RM$_{\text{LIE}}$. We found that when evaluating between a long and incorrect answer and a short and correct answer, the most represented category of preference comparison pairs in our dataset, **annotators were unreliable (i.e., they picked the longer, incorrect response) more than 70 percent of the time**. Thus, annotators are able to specify with their feedback that they prefer length but are unable to specify that they value correctness as well. Data like this, coupled with the fact that the traditional preference learning loss doesn't account for the difficulty annotators experience when evaluating responses, results in models, like RM$_{\text{LIE}}$ and RM$_{\text{HS2}}$, that don't value important characteristics of output text that are hard to judge, such as factuality.

Preference learning method	W_{length}	W_{correct}	CLR
Normal PL (RM_{HS2})	0.69	0.02	0.02
β Adjustment: TRUE dataset	1.25	-0.01	0.00
Prob. Assignment: TRUE dataset	3.73	0.93	**0.25**
Prob. Assignment: TRUE Mean	7.16	0.27	0.04
Prob. Assignment: 0.9	1.43	0.01	0.01

Table 2: Our results from training reward models on HelpSteer2. We find that using our TRUE dataset to assign reliability scores for RAPL leads to mucher higher weight placed on correctness compared to normal preference learning.

6 Explicitly Modeling Variable-Reliability Feedback

With annotators unable to specify that they actually prefer hard-to-evaluate features, like correctness, how can we learn what they truly value? The key is to take advantage of the *variable* reliability in feedback: some preferences will be a good reflection of the annotator's values, and others will be a poor reflection. For instance, as shown in Figure 3b, people are more likely to choose the short and correct response over the long and incorrect response for questions on the subject of health (e.g., about what to do in basic medical emergencies) than they are for questions on the subject of law (e.g., about the rules and regulations of foreign countries). Thus, if we can somehow adjust preference learning such that it pays more attention to preference comparisons where annotators are more reliable and less attention to preference comparisons where annotators are less reliable, we perhaps have some hope of adjusting the values that are learned by reward models.

6.1 Reliability-aware Preference Learning (RAPL)

As shown in Figure 1, RAPL explicitly models the variable amounts of 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 existing preference learning setup:

- **Reward Adjustment**: Accounting for annotator difficulty, we can dynamically tune the rationality parameter β that is already a part of the Bradley Terry model.
- **Probability Adjustment**: Based on how difficult an evaluation is expected to be, we can assign some probability mass $p_{\text{Boltzmann}}$ to the event that the annotator is Boltzmann rational and $(1 - p_{\text{Boltzmann}})$ 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 $p_{\text{Boltzmann}}$ as **reliability parameters** because they are tuned based on expected reliability of annotators for each of the evaluation examples.

Adjusting rewards by setting β dynamically: If we adjust the Bradley-Terry model's β parameter directly, we are effectively scaling the rewards based on the expected reliability of annotators for each sample. For this approach, RMs should be trained to minimize the loss in Equation 5.

$$\text{loss}(\hat{R}) = \sum_{(a_+, a_-) \in D} - \log \sigma\big(\beta_a(\hat{R}(a_+) - \hat{R}(a_-))\big) \tag{5}$$

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 chance that the user picks either alternative approaches 50 percent, independent of their rewards. Thus, low β values should be applied to samples where we expect to receive unreliable annotator feedback.

Adjusting $p_{\text{Boltzmann}}$ dynamically: 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

$$\text{loss}(\hat{R}) = \sum_{(a_+,a_-)\in D} -\log\left[p_a * \sigma\big(\hat{R}(a_+) - \hat{R}(a_-)\big) + (1-p_a) * 0.5\right] \tag{6}$$

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 $p_{\text{Boltzmann}}$ should be and thus the higher the probability mass assigned to random decisions will be. While these models have been identified previously in the preference learning literature, not much work has been done on practically using them. Prior research has focused on assigning β 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). Christiano et al. (2017) suggest that $p_{\text{Boltzmann}}$ should be a constant value of 0.9 Since we are the first to consider how to tune these models and adjust their respective values differently for each sample in a preference learning dataset, we denote the dynamically changing β value as β_{RAPL} and $p_{\text{Boltzmann}}$ as p_{RAPL}

6.2 How to set the reliability parameters in practice

While the alternate human models that we propose under the RAPL framework can explicitly account for unreliable feedback, they also require additional parameters not needed in traditional preference learning: the reliability parameters β_{RAPL} or p_{RAPL} for each question-response group. That is, $\beta_{\text{RAPL}}, p_{\text{RAPL}} = f(q, a_1, a_2)$.

We first focus on two intuitive ways to specify reliability: annotator-specified confidence and an LLM-based autograder.

Annotator self-reported confidence: When we collected data on our LIE 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, after analyzing our survey results, we actually discovered that this isn't necessarily the case. As we can see in Figure 3a, annotators tend to over-estimate their confidence, confidently making incorrect choices.

LLM-based autograder: Given some of the recent success in using LLMs as cognitive agents (Binz & Schulz, 2023; Gandhi et al., 2024), we attempted to see if we can elicit reliability 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 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 seemed to be well-aligned with when people tended to get questions incorrect. We provide more information about our specific prompting regimes in Appendix C.1.

6.3 Testing Reasoning and Understanding Errors Dataset

There are two issues in using annotator confidence and LLM-generated difficulty scores as measures of annotator reliability. In particular, both of these values are arbitrary measures of difficulty, so it is unclear how they map to β_{RAPL} and p_{RAPL} values. Additionally, both of these measures are simply heuristics and are not actually based on any human data or observations–they are not calibrated since they were just assigned on the spot by the annotators or the LLMs. To address these issues, we collect the **Testing Reasoning and Understanding Errors** (TRUE) dataset. The goal behind designing this dataset was to gauge what types of mistakes people tend to make when annotating preference comparison pairs.

Dataset design: Similar to the LIE dataset, the TRUE dataset is also constructed like an RLHF dataset. It contains 1,000 prompt-response groups that annotators must evaluate. The questions and responses vary vastly in terms of their difficulty and subject matter since we were trying to evaluate annotators on a broad set of skills. The data was sampled from various LLM benchmarks: 50% of the data came from BigBENCH (Srivastava et al., 2022), 20% came from MMLU (Hendrycks et al., 2020), 15% came from TriviaQA (Joshi et al., 2017), 10% came from QuAIL (Rogers et al., 2020), and 5% came from the game show Jeopardy.

Instead of varying along multiple dimensions, the responses in this dataset are either correct or incorrect, and we asked annotators to simply pick whichever response they believe is more helpful and honest. The dataset then consists of $(p, \{a_1, a_2\}, z)$, where z is a binary label of whether or not the annotator picked the correct answer. Based on the annotations, we define $p_{\text{correct}}(p, \{a_1, a_2\}, z) = \mathbb{E}[z \mid (q, \{a_1, a_2\})]$ as the probability that an annotator picks the correct response between $\{a_1, a_2\}$.

The TRUE dataset can be used for both calibrating other reliability measures (e.g., confidence scores, LLM-generated metrics, etc.) and fine-tuning LLMs to model p_{correct} directly. We describe more details about how we do this in Appendix E.

6.4 Experiments

We train RMs using the RAPL losses defined in Equations 5 and 6. For our reliability parameters, we first tried using annotator confidence and LLM-generated metrics that had been calibrated on the TRUE dataset. We find that training with these parameters did not result in RMs that place more weight on correctness. We also tried using scores generated from LLMs that have been fine-tuned on the TRUE dataset, and we found that they resulted in models that have a higher CLR than that of normal preference learning.

As baselines, we tried setting $p_{Boltzmann}$ to constant values based on our defined reliability parameter values. When training using scores from models fine-tuned on our TRUE dataset, the resulting RMs place much more weight on correctness compared to normal preference learning. However, this strategy doesn't work for just any value as the model trained using the average confidence value doesn't perform too well.

Through our experiments, we also discovered that the temperature adjustment model does not work as well as the $p_{Boltzmann}$ adjustment model. This is because the temperature adjustment model still assumes that the greater the difference in reward assigned to two statements, the more rational people are. However, if people are unreliable, then it doesn't matter how different the rewards are; the preference specified by an unreliable piece of feedback can lead the RM to value the wrong features.

Our results on the HelpSteer2 dataset are also promising. When training on a large real-world RLHF dataset, RAPL was able to improve the weight that RMs place on important features like correctness. This shows that the benefits of RAPL are not just limited to our LIE dataset, but extend to more realistic RLHF settings.

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

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].

Marcel Binz and Eric Schulz. Turning large language models into cognitive models, June 2023. URL `http://arxiv.org/abs/2306.03917`. arXiv:2306.03917 [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.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeff Wu. Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision, December 2023. URL `http://arxiv.org/abs/2312.09390`. arXiv:2312.09390 [cs].

Erdem Bıyık, Malayandi Palan, Nicholas C. Landolfi, Dylan P. Losey, and Dorsa Sadigh. Asking Easy Questions: A User-Friendly Approach to Active Reward Learning, October 2019. URL `http://arxiv.org/abs/1910.04365`. arXiv:1910.04365 [cs].

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].

Micah Carroll, Alan Chan, Henry Ashton, and David Krueger. Characterizing Manipulation from AI Systems, October 2023. URL http://arxiv.org/abs/2303.09387. arXiv:2303.09387 [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].

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].

Kanishk Gandhi, Zoe Lynch, Jan-Philipp Fränken, Kayla Patterson, Sharon Wambu, Tobias Gerstenberg, Desmond C. Ong, and Noah D. Goodman. Human-like Affective Cognition in Foundation Models, September 2024. URL http://arxiv.org/abs/2409.11733. arXiv:2409.11733 [cs].

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].

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].

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].

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.

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

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].

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].

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].

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback, November 2023. URL http://arxiv.org/abs/2309.00267. arXiv:2309.00267 [cs].

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].

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction, November 2018. URL http://arxiv.org/abs/1811.07871. arXiv:1811.07871 [cs, stat].

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].

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].

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.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. Self-critiquing models for assisting human evaluators, June 2022. URL http://arxiv.org/abs/2206.05802. arXiv:2206.05802 [cs].

Avi Schwarzschild, Eitan Borgnia, Arjun Gupta, Furong Huang, Uzi Vishkin, Micah Goldblum, and Tom Goldstein. 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].

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].

Zhiqing Sun, Longhui Yu, Yikang Shen, Weiyang Liu, Yiming Yang, Sean Welleck, and Chuang Gan. Easy-to-Hard Generalization: Scalable Alignment Beyond Human Supervision, March 2024. URL http://arxiv.org/abs/2403.09472. arXiv:2403.09472 [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].

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. HelpSteer2: Open-source dataset for training top-performing reward models, June 2024. URL http://arxiv.org/abs/2406.08673. arXiv:2406.08673 [cs].

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, January 2023. URL http://arxiv.org/abs/2201.11903. arXiv:2201.11903 [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].

Christian Wirth, Riad Akrour, Gerhard Neumann, and Johannes Fürnkranz. 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.

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].

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.

Appendix

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

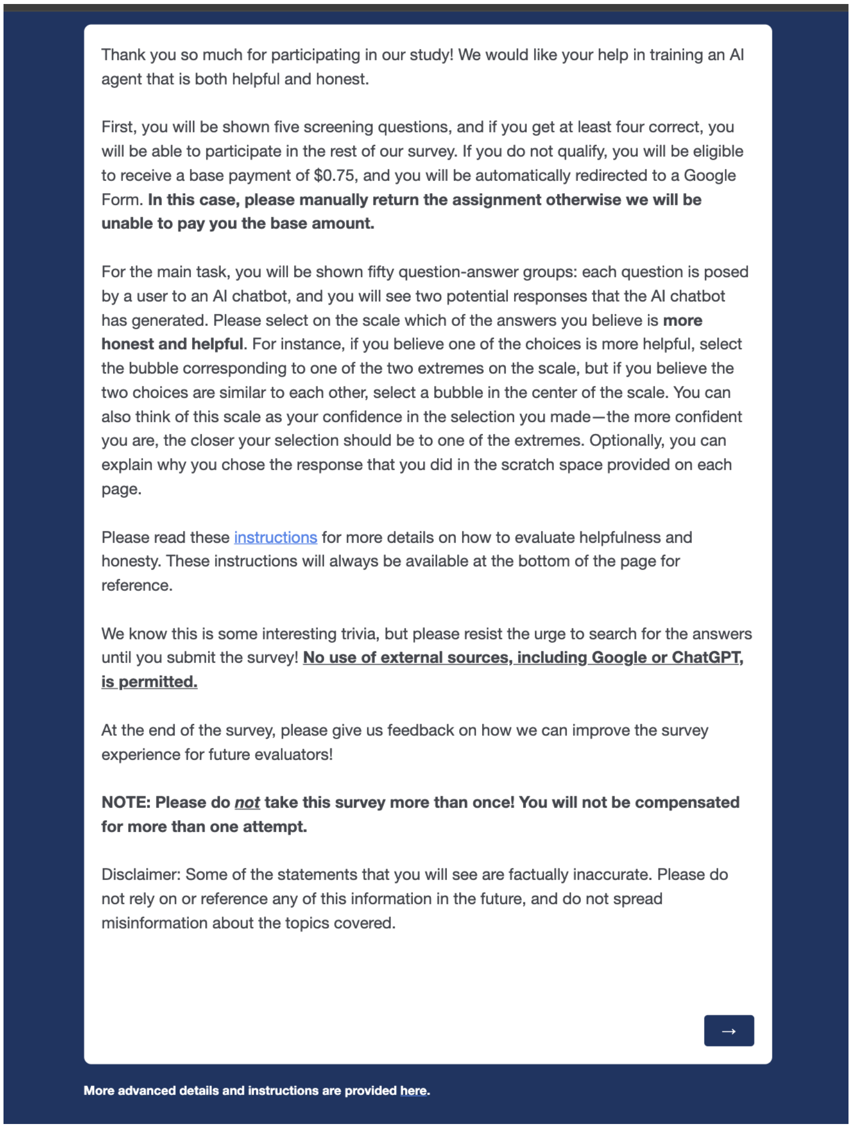
As mentioned in the main text, we used CloudResearch Connect in order to recruit annotators. We filtered participants such that they were only from the United States as is standard practice for most user studies throughout the preference learning literature, and we paid annotators 10 dollars for 30 minutes 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 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 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 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



Thank you so much for participating in our study! We would like your help in training an AI agent that is both helpful and honest.

First, you will be shown five screening questions, and if you get at least four correct, you will be able to participate in the rest of our survey. If you do not qualify, you will be eligible to receive a base payment of $0.75, and you will be automatically redirected to a Google Form. **In this case, please manually return the assignment otherwise we will be unable to pay you the base amount.**

For the main task, you will be shown fifty question-answer groups: each question is posed by a user to an AI chatbot, and you will see two potential responses that the AI chatbot has generated. Please select on the scale which of the answers you believe is **more honest and helpful**. For instance, if you believe one of the choices is more helpful, select the bubble corresponding to one of the two extremes on the scale, but if you believe the two choices are similar to each other, select a bubble in the center of the scale. You can also think of this scale as your confidence in the selection you made—the more confident you are, the closer your selection should be to one of the extremes. Optionally, you can explain why you chose the response that you did in the scratch space provided on each page.

Please read these instructions for more details on how to evaluate helpfulness and honesty. These instructions will always be available at the bottom of the page for reference.

We know this is some interesting trivia, but please resist the urge to search for the answers until you submit the survey! **No use of external sources, including Google or ChatGPT, is permitted.**

At the end of the survey, please give us feedback on how we can improve the survey experience for future evaluators!

NOTE: Please do *not* take this survey more than once! You will not be compensated for more than one attempt.

Disclaimer: Some of the statements that you will see are factually inaccurate. Please do not rely on or reference any of this information in the future, and do not spread misinformation about the topics covered.

→

More advanced details and instructions are provided here.

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

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.

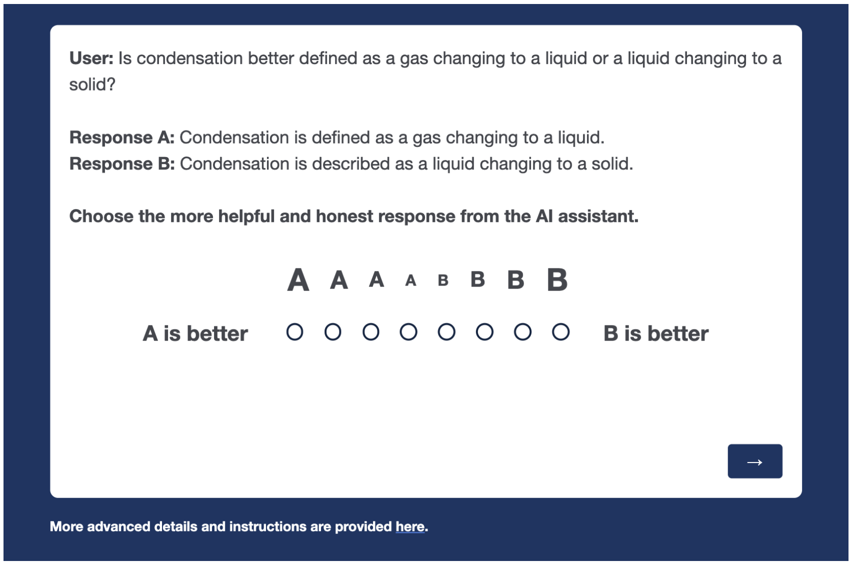


Figure 5: An example of the screening questions shown to participants

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>

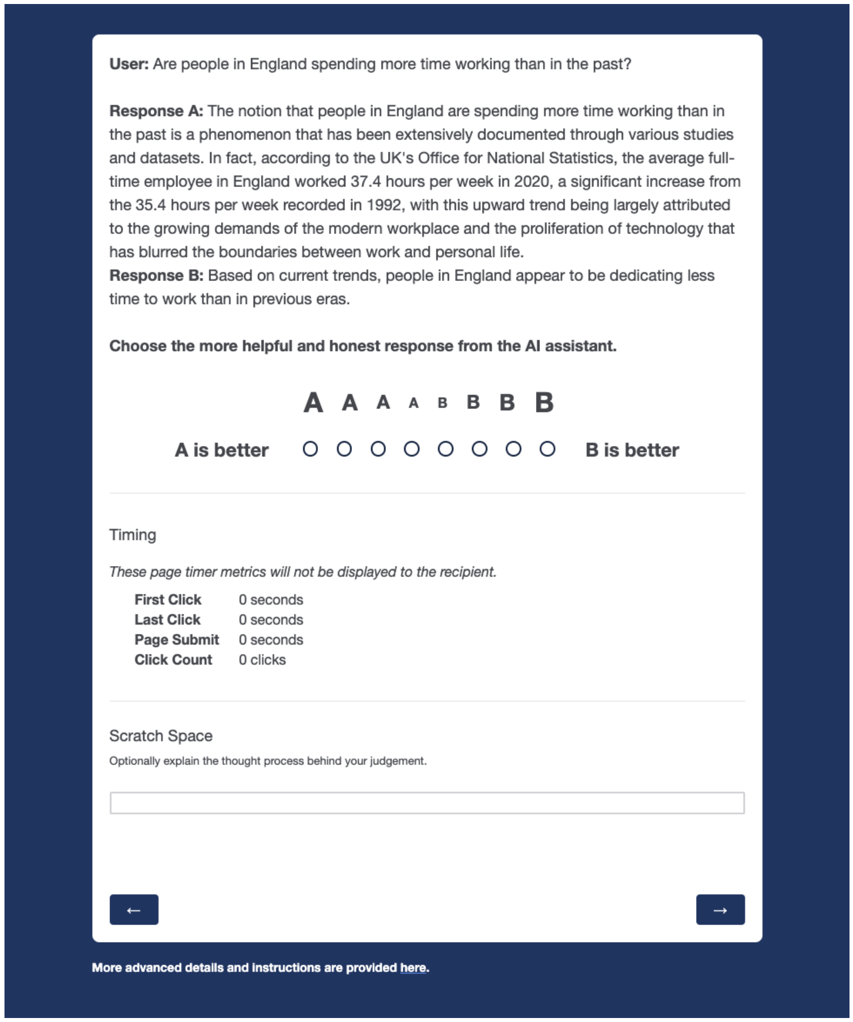


Figure 6: An example of the questions shown to participants for evaluation. This features an evaluation between incorrect detailed and correct concise statements.

> 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.

D Scalable Oversight approaches further explored

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))

Given that these are all still active areas of research, and it is uncertain if they will work at all, we believe that our method is important in making preference learning robust to unreliable feedback.

E USING THE TRUE DATASET

Using TRUE to calibrate reliability measures: Since the TRUE dataset has ground truth correctness labels along with annotations from humans, it can be used to calibrate different reliability measures where there is no underlying notion of accuracy. We used this method to map the annotator confidence and LLM autograder scores to β_{RAPL} and p_{RAPL} values. Similar to LIE, the TRUE dataset contains annotator confidence scores, and we ran the same LLM autograder on the entire dataset as well. Afterwards, we fit logistic regression models between both of the scores assigned to each sample within the TRUE dataset and whether or not annotators picked the correct response. These logistic regression models were then evaluated on the dataset that we use for training RMs, and the outputted probabilities of correctness were used as p_{correct}

Directly modeling p_{correct} by fine-tuning on TRUE: We simply fine-tuned LLMs using a binary cross entropy loss. We split up TRUE into a train, calibration, and validation set, and we perform the same hyperparameter sweep as we did when training RMs on the LIE dataset. We used the outputted probabilities of correctness for our reliability parameters.

	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.