REASONS TO REJECT? ALIGNING LANGUAGE MODELS WITH JUDGMENTS

Weiwen Xu[♡♠∗] **Deng Cai**[♡†] **Zhisong Zhang**[♡] **Wai Lam**[♠] **Shuming Shi**[♡]
[♡]Tencent AI Lab [♠]The Chinese University of Hong Kong
{wwxu,wlam}@se.cuhk.edu.hk
{jcykcai,zhisonzhang,shumingshi}@tencent.com

ABSTRACT

As humans, we consistently engage in interactions with our peers and receive feedback in the form of natural language. This language feedback allows us to reflect on our actions, maintain appropriate behavior, and rectify our errors. The question arises naturally: can we use language feedback to align large language models (LLMs)? In contrast to previous research that aligns LLMs with reward or preference data, we present the first systematic exploration of alignment through the lens of language feedback (i.e., judgment). We commence with an in-depth investigation of potential methods that can be adapted for aligning LLMs with judgments, revealing that these methods are unable to fully capitalize on the judgments. To facilitate more effective utilization of judgments, we propose a novel framework, Contrastive Unlikelihood Training (CUT), that allows for fine-grained inappropriate content detection and correction based on judgments. Our offline alignment results show that, with merely 1317 off-the-shelf judgment data, CUT (LLaMA2-13b) can beat the 175B DaVinci003 and surpass the best baseline by 52.34 points on AlpacaEval. The online alignment results demonstrate that CUT can align LLMs (LLaMA2-chat-13b) in an iterative fashion using model-specific judgment data, with a steady performance improvement from 81.09 to 91.36 points on AlpacaEval. Our analysis further suggests that judgments exhibit greater potential than rewards for LLM alignment and warrant future research.[1]

1 INTRODUCTION

Large language models (LLMs) acquire extensive knowledge and remarkable reasoning capabilities through self-supervised pre-training on large-scale corpora (Brown et al., 2020; Du et al., 2022; Touvron et al., 2023). To unleash the power of pre-trained LLMs for addressing real-world applications, it is crucial to ensure LLMs can follow human intentions and values (Ouyang et al., 2022). This process, known as alignment, has the potential to pave the way for a future in which artificial intelligence (AI) serves as a helpful and reliable partner for humanity (Wang et al., 2023b).

Figure 1 shows three typical paradigms to achieve alignment. The most straightforward one is learning from demonstrations, wherein human demonstrations of desired responses to a set of instructions are collected to fine-tune LLMs using supervised learning (Wei et al., 2022; Ouyang et al., 2022). However, the performance gains diminish rapidly when scaling up the data size (Zhou et al., 2023). On the other hand, learning from feedback offers a more scalable approach (Ouyang et al., 2022; Glaese et al., 2022; Bai et al., 2022a). One key advantage of feedback over demonstrations is that feedback can convey both positive and negative aspects, enabling the model to discern desirable and undesirable outcomes. In addition, feedback is usually tailored to the current model-generated responses, adhering to the principle of teaching according to the learner's aptitude.

Prior research on learning from feedback has been primarily focused on incorporating value feedback (i.e., scalar rewards) into the alignment process. The most widely adopted reinforcement learning (RL)

[∗]Work done during an internship at Tencent AI Lab.
[†]Corresponding author.
[1]Our source codes and models are publicly available at https://github.com/wwxu21/CUT

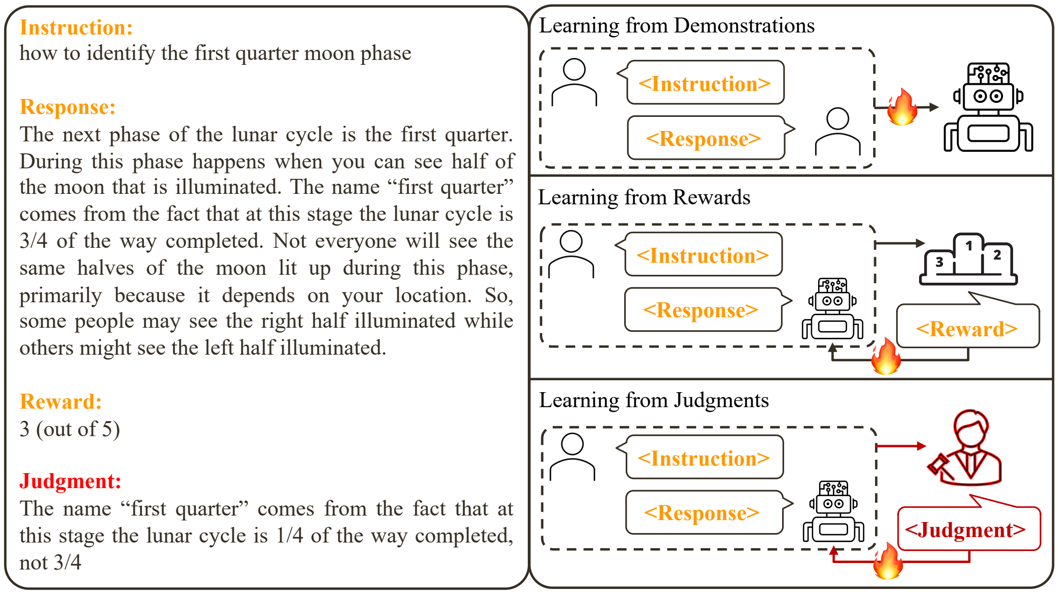


Figure 1: The illustration of three paradigms for aligning LLMs.

techniques, particularly proximal policy optimization (PPO) (Schulman et al., 2017; Ouyang et al., 2022), optimize an LLM to maximize the scalar rewards of its generated responses. Nevertheless, PPO is known to be complex and often unstable (Zheng et al., 2023), which has prompted numerous efforts to simplify or stabilize the training process (Ramamurthy et al., 2023; Peng et al., 2023b; Dong et al., 2023; Touvron et al., 2023). Another strand of work, referred to as Hindsight (Zhang et al., 2023; Liu et al., 2023a), transforms scalar rewards to language instructions and employs supervised training on the updated data.

Language feedback (i.e., judgment) is another kind of feedback that provides more nuanced commendations and critiques through natural language expressions. Unlike scalar rewards, which are information-sparse for solely indicating the goodness of a response, judgments can elucidate the specific aspects that are good or bad, the rationale behind their evaluation, and suggestions for improvement. The above advantages suggest that aligning LLMs with judgments can be more efficient and effective (Saunders et al., 2022). However, current approaches merely use judgments to prompt LLMs for an improved response, which is subsequently employed as a new demonstration for supervised training (Bai et al., 2022b; Scheurer et al., 2022; 2023). This indirect utilization of judgments suffers from the incapability to learn from mistakes, which is the core spirit of learning from feedback, and is constrained by the refinement capabilities of LLMs.

In this study, we present an extensive investigation of potential methods that can be adapted for *aligning LLMs with judgments*. To facilitate a comprehensive aligning process, we propose a novel framework, Contrastive Unlikelihood Training (CUT), that enables fine-grained inappropriate content detection and correction based on judgments. The core idea of CUT is to detect and penalize inappropriate content in a response by contrasting its generation probabilities guided by an authentic judgment that may contain negative opinions and a fabricated judgment portraying perfection.

We carry out alignment experiments in both offline and online settings, wherein the target LLM learns from the off-the-shelf judgments and the judgments derived from self-generated responses, respectively. Extensive results on offline alignment demonstrate the effectiveness of CUT in learning from judgments in both cold-start (using unaligned base models such as LLaMA2) and warm-start (using aligned base models such as LLaMA2-chat) scenarios. Notably, when trained with only 1317 offline judgment data, CUT (LLaMA2-13b) attains a winning rate of 62.56 (beats the 175B DaVinci003[2]) and outperforms the best baseline by 52.34 points on AlpacaEval. Furthermore, our online alignment experiments show that CUT is capable of iteratively refining LLMs with up-to-date, model-specific judgments. For example, we observe a consistent performance improvement on

[2]https://platform.openai.com/docs/model-index-for-researchers

LLaMA2-chat-13b over four times of CUT iterations, rising from 81.09 to 91.36 points on AlpacaEval. Our analysis comparing rewards and judgments suggests that aligning LLMs with judgments offers significant potential and warrants future research. Our contributions can be summarized as follows.

- We present the first systematic exploration of aligning LLMs with judgments.
- We introduce a novel framework, CUT, that facilitates the alignment of LLMs through direct and explicit learning from judgments. Notably, CUT allows fine-grained inappropriate content detection and correction based on judgments.
- Our results showcase the effectiveness of CUT in aligning LLMs across cold-start and warm-start scenarios, generalist and specialist applications, as well as offline and online settings.
- Our analysis indicates that judgments hold promising potential over rewards for aligning LLMs.

2 Related Work

2.1 Collecting Feedback

Value Feedback (Reward). Traditional RL research for natural language processing (NLP) uses algorithmically defined metrics as reward functions, such as BLEU for translation (Wu et al., 2016) and ROUGE for summarization (Ranzato et al., 2016). For LLM alignment, existing works primarily leverage human preference data to fit a reward model, which subsequently generates scalar rewards (Ouyang et al., 2022). To augment the informativeness of value feedback, recent studies introduce rewards for multiple dimensions (Bai et al., 2022a; Touvron et al., 2023; Wu et al., 2023) and provide rewards for each sub-step (Lightman et al., 2023).

Language Feedback (Judgment). Judgments typically necessitate human annotations on the model-generated responses. There are several works where judgments are collected for specific tasks, such as dialogue (Xu et al., 2023b), summarization (Saunders et al., 2022; Scheurer et al., 2022; 2023; Liu et al., 2023c), question answering (Li et al., 2022; Xu et al., 2023a), script generation (Tandon et al., 2022), and general instruction-following tasks (Wang et al., 2023a). Another direction is to train an AI judge to automatically provide precise judgments for the model's responses (Bai et al., 2022b; Akyurek et al., 2023; Li et al., 2023).

2.2 Learning from Feedback

Existing approaches for learning from feedback can be classified into two distinct categories: prompting and fine-tuning, differentiated by whether updates to the LLMs' parameters are absent or present.

Prompting. Prompting does not alter the parameters of the LLMs. Instead, it leverages language feedback on previous responses to prompt the generation of improved responses (Welleck et al., 2022; Akyurek et al., 2023). The language feedback can be sourced from diverse aspects Nathani et al. (2023); Yu et al. (2023) and the refinement process can be iterated multiple times Yang et al. (2022); Peng et al. (2023a); Madaan et al. (2023). However, these approaches consume more computation than single-pass generation and usually rely on the in-context learning capabilities of the LLMs (Brown et al., 2020; Liu et al., 2023b).

Fine-tuning. Fine-tuning aims to directly train a better LLM. In this context, value feedback has been extensively used through RL, particularly PPO (Schulman et al., 2017; Ziegler et al., 2019; Stiennon et al., 2020; Ouyang et al., 2022; Bai et al., 2022a; Yang et al., 2023). However, these RL approaches are notoriously unstable and complex (Zheng et al., 2023). To stabilize RL, Ramamurthy et al. (2023) propose to reduce the action space through truncation and Peng et al. (2023b) employ an advantage model and selective rehearsal. In addition, many efforts have been put into designing simpler alternatives to RL. Dong et al. (2023); Touvron et al. (2023) treat value feedback as a ranking criterion and simply train models using the best model-generated responses. There are also attempts to leverage the results of prompting for training a better model. That is, the improved response elicited by language feedback is employed as new training data (Bai et al., 2022b; Scheurer et al., 2022;

2023). However, these methods still suffer from the incapability to learn from mistakes. Rafailov et al. (2023); Yuan et al. (2023); Song et al. (2023) demonstrate that LLMs themselves can be used as reward functions and derive different training objectives to eliminate the need for RL. Zhang et al. (2023); Liu et al. (2023a) relabel the input using the value feedback received by the response, referred to as *Hindsight*. This hindsight method allows LLMs to learn to generate responses of different qualities. In this work, our CUT is a novel fine-tuning method that allows LLMs to comprehensively learn from both positive and negative aspects of language feedback.

3 Preliminaries

In this section, we first lay out a formal problem definition of *aligning LLMs with judgments* and then present a survey of three potentially useful methods that can be adapted for tackling this problem.

3.1 Problem Setting

Suppose that there is a set of instruction-response-judgment triplets $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{j})$, where the instruction $\boldsymbol{x} = [x_1, \ldots, x_M]$, the response $\boldsymbol{y} = [y_1, \ldots, y_N]$, and the judgment $\boldsymbol{j} = [j_1, \ldots, j_Q]$ are token sequences of length M, N, and Q, respectively. The response may exhibit certain flaws or be considered entirely satisfactory. The judgment provides an analysis of the strengths and weaknesses of the response. The judgment can be drafted either by human annotators[3] or AI judge models (Akyurek et al., 2023; Li et al., 2023). The goal of aligning LLMs with judgments is to enable the LLM to retain appropriate behaviors mentioned in the strengths, and more importantly, address the weaknesses to prevent future misbehavior.

Depending on whether the responses $\boldsymbol{y}$ are generated from the LLM to be aligned, the learning process can be classified into two distinct types: *offline alignment* and *online alignment*. In offline alignment, the target LLM learns from a static, off-the-shelf, model-agnostic dataset. Conversely, in online alignment, the target LLM reflects on its own outputs through direct interactions with a judge. This online alignment process can be conducted iteratively, akin to how humans continuously improve their skills by receiving ongoing feedback from others over time.

3.2 Potential Solutions

Forward Prediction. Forward prediction refers to the process of sequentially predicting the response and its judgment, which was originally proposed in the context of dialogue generation (Weston, 2016; Li et al., 2017). It can be seamlessly adapted to the alignment of LLMs. Specifically, the LLM is trained under the maximum likelihood estimation (MLE) objective to first generate the response $\boldsymbol{y}$ based on the instruction $\boldsymbol{x}$ and subsequently generate the judgment $\boldsymbol{j}$ based on the combined sequence $[\boldsymbol{x}, \boldsymbol{y}]$.

$$L^f(\boldsymbol{x}, \boldsymbol{j}, \boldsymbol{y}) = -\frac{1}{N}\sum_t \log p_\theta(y_t|y_{<t}, \boldsymbol{x}) - \frac{1}{Q}\sum_t \log p_\theta(j_t|j_{<t}, \boldsymbol{y}, \boldsymbol{x}) \tag{1}$$

where θ represents the trainable parameters of the target LLM.

Imitation Learning from Language Feedback. Imitation learning from Language Feedback (ILF) asks the LLM to refine the initial response $\boldsymbol{y}$ given the feedback $\boldsymbol{j}$ (Bai et al., 2022b; Scheurer et al., 2022; 2023). The improved response $\hat{\boldsymbol{y}}$, paired with the initial instruction $\boldsymbol{x}$, is used to fine-tune the LLM under the MLE objective.

$$\hat{\boldsymbol{y}} = \mathbf{LLM}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{j}) \tag{2}$$

$$L^i(\boldsymbol{x}, \hat{\boldsymbol{y}}) = -\frac{1}{N}\sum_t \log p_\theta(\hat{y}_t|\hat{y}_{<t}, \boldsymbol{x}) \tag{3}$$

[3]We argue that annotating judgments is not more difficult than annotating scalar rewards (or preferences), as annotators typically assign scalar rewards based on specific reasons. Essentially, we are just asking annotators to write down the reasons behind their decisions.

	Instruction: $\boldsymbol{x}$	Response: $\boldsymbol{y}$	Judgment: $\boldsymbol{j}$	Aligned $\boldsymbol{x} \rightarrow \boldsymbol{y}$	Aligned $[\boldsymbol{x}, \boldsymbol{j}] \rightarrow \boldsymbol{y}$
Align-P	James buys 5 packs of beef that are 4 pounds each. The price of beef is $5.50 per pound. How much did he pay?	He bought 5 * 4 = 20 pounds of beef. So he paid 20 * 5.5 = $110.	Your response to the instruction is satisfactory.	✓	✓
Align-N	James buys 5 packs of beef that are 4 pounds each. The price of beef is $5.50 per pound. How much did he pay?	James had 4 packs of beef that were 5 pounds each. Each pack was 5 pounds and it cost 5.50. So 5 * 5.50 = 27.50 dollars.	The answer forgets to multiply the total amount of pounds of beef (5*4).	✗	✓
Misalign	James buys 5 packs of beef that are 4 pounds each. The price of beef is $5.50 per pound. How much did he pay?	James had 4 packs of beef that were 5 pounds each. Each pack was 5 pounds and it cost 5.50. So 5 * 5.50 = 27.50 dollars.	Your response to the instruction is satisfactory.	✗	✗

Table 1: The illustration of three categories of alignment data. The "Aligned" columns indicate if the response aligns with the instruction or the combination of instruction and judgment, respectively.

Hindsight. Hindsight methods (Zhang et al., 2023; Liu et al., 2023a) rewrite the instruction $\boldsymbol{x}$ based on the scalar rewards received by the response $\boldsymbol{y}$. For instance, if a response receives a scalar reward below a certain threshold, the phrase "generate a correct answer" is appended to the original instruction; otherwise, "generate an incorrect answer" is added. Obviously, this approach can be naturally extended to our problem setting. Concretely, the LLM is trained to generate the response $\boldsymbol{y}$ conditioned on the sequence $[\boldsymbol{x}, \boldsymbol{j}]$.

$$L^h(\boldsymbol{x}, \boldsymbol{j}, \boldsymbol{y}) = -\frac{1}{N} \sum_t \log p_\theta(y_t | y_{<t}, \boldsymbol{x}, \boldsymbol{j}) \tag{4}$$

However, in forward prediction, learning to generate judgments does not necessarily translate into enhanced response generation, given that response generation precedes judgment generation. ILF only makes use of the positive data (i.e., the improved responses), limiting the model's capacity to recognize and rectify weaknesses or errors underscored in negative judgments. As for Hindsight, employing unsatisfactory responses as MLE targets inevitably increases the risk of generating unsatisfactory responses. In summary, we contend that existing methods cannot take full advantage of judgments, which motivates us to design a better solution.

4 Contrastive Unlikelihood Training

To overcome the limitations mentioned in § 3, we propose Contrastive Unlikelihood Training (CUT), a novel fine-tuning framework to align LLMs with judgments. The central idea of CUT can be summarized as **Learning from Contrasting**. We contrast the response generation under different conditions to shed light on the appropriate behavior that the LLM should maintain, as well as the specific content necessitating adjustments. Based on these insights, we use MLE training for appropriate content and Unlikelihood Training (UT) for inappropriate content.

4.1 Incorporating Judgments for Alignment

We call an instruction-response pair "aligned" if the response follows the instruction faithfully and satisfies human expectations $\boldsymbol{x} \rightarrow \boldsymbol{y}$. Otherwise, a judgment describes the errors or deficiencies present in the response. Assuming the task is to generate a response that intentionally fulfills the judgment, it can be inferred that the response always aligns with the combined input of instruction and judgment $[\boldsymbol{x}, \boldsymbol{j}] \rightarrow \boldsymbol{y}$. Based on the idea, we construct three types of alignment data, depicted in Table 1.

Align-P: The LLM produces a satisfactory response $\boldsymbol{y}$ to the original instruction $\boldsymbol{x}$. Therefore, a positive judgment $\boldsymbol{j}$ is conferred to acknowledge the commendable performance of the LLM. It is evident that the response $\boldsymbol{y}$ is aligned with the instruction $\boldsymbol{x}$ as well as the combined input $[\boldsymbol{x}, \boldsymbol{j}]$.

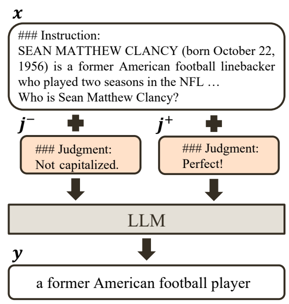
Align-N: The LLM makes some mistakes in its generation, resulting in an unsatisfactory response $\boldsymbol{y}$. Consequently, a negative judgment $\boldsymbol{j}$ details the corresponding critiques. For Align-N, $\boldsymbol{y}$ is not aligned with original instruction $\boldsymbol{x}$. However, when considering $\boldsymbol{x}$ and $\boldsymbol{j}$ as a whole, $\boldsymbol{y}$ is indeed aligned with the combined input $[\boldsymbol{x}, \boldsymbol{j}]$.

Misalign: The authentic negative judgment in Align-N is substituted with a fake positive judgment $\boldsymbol{j}$. In this case, the response $\boldsymbol{y}$ is not aligned with either the original instruction $\boldsymbol{x}$ or the combination of instruction and judgment $[\boldsymbol{x}, \boldsymbol{j}]$.

4.2 Learning from Contrasting

With the above three categories of alignment data. We can deduce two notable contrasts that provide valuable insights to guide the alignment of LLMs.

Align-N vs. Misalign: Despite Align-N and Misalign are not aligned in terms of $\boldsymbol{x} \rightarrow \boldsymbol{y}$, they show opposite polarities in the task of $[\boldsymbol{x}, \boldsymbol{j}] \rightarrow \boldsymbol{y}$. Thanks to the strong in-context learning capabilities of LLMs, the alignment flip from Align-N (aligned) to Misalign (misaligned) is often accompanied by decreased generation probabilities of the response, particularly for tokens that exhibit a strong correlation with the authentic negative judgment. Figure 2 presents a simple example, wherein the response commits a minor capitalization issue. The LLM assigns a considerably higher probability for "a" when taking the authentic negative judgment $\boldsymbol{j}^-$ instead of the fake positive judgment $\boldsymbol{j}^+$ as additional input, precisely at the point where the LLM commits the error.



y	$p_\theta(y\|x, j^-)$	$p_\theta(y\|x, j^+)$	Objective
a	**0.82**	**0.03**	**UT**
former	0.94	0.94	MLE
American	0.76	0.93	MLE
football	0.99	0.99	MLE
player	0.02	0.03	MLE
</s>	0.89	0.85	MLE

Figure 2: The probability of generating identical output text in an LLM under Align-N (left) and Misalign (right) contexts.

To take advantage of the above contrast, we feed Align-N and Misalign examples to the LLM to get token generation probabilities $p_\theta(y_t|\boldsymbol{y}_{<t}, \boldsymbol{x}, \boldsymbol{j}^-)$ and $p_\theta(y_t|\boldsymbol{y}_{<t}, \boldsymbol{x}, \boldsymbol{j}^+)$ separately. We consider the tokens that display a substantially increased generation probability when conditioned on $\boldsymbol{j}^-$ compared to $\boldsymbol{j}^+$ as inappropriate tokens (e.g., "a" in Figure 2). Concretely, the following criterion is adopted:

$$U = \{t \mid p_\theta(y_t|\boldsymbol{y}_{<t}, \boldsymbol{x}, \boldsymbol{j}^-) - \lambda \cdot p_\theta(y_t|\boldsymbol{y}_{<t}, \boldsymbol{x}, \boldsymbol{j}^+) > 0\} \tag{5}$$

where $\lambda \geq 1$ is a hyperparameter to tradeoff the precision and recall of detecting inappropriate tokens.

We apply the UT objective (Welleck et al., 2020) on the identified inappropriate tokens for pushing the LLM to explore alternative generations. For other tokens, we use the standard MLE loss:

$$L_1 = -\frac{1}{N}\Big(\sum_{t \notin U} \log p_\theta(y_t|\boldsymbol{y}_{<t}, \boldsymbol{x}) + \sum_{t \in U} \alpha \log(1 - p_\theta(y_t|\boldsymbol{y}_{<t}, \boldsymbol{x}))\Big) \tag{6}$$

where α is another hyperparameter to control the scale of unlikelihood loss.

Align-P vs. Align-N: Despite both Align-P and Align-N are aligned in terms of $[\boldsymbol{x}, \boldsymbol{j}] \rightarrow \boldsymbol{y}$, only Align-P is aligned when solely considering the instruction ($\boldsymbol{x} \rightarrow \boldsymbol{y}$). Essentially, it suggests that the LLM should output different responses depending on whether a negative judgment is incorporated or not. Therefore, the comparison provides valuable information for the LLM to discern satisfactory and unsatisfactory responses. Specifically, we train on this comparison with the following MLE objective:

$$L_2 = -\frac{\mathbb{1}(\boldsymbol{x} \rightarrow \boldsymbol{y})}{N} \sum_t \log p_\theta(y_t|\boldsymbol{y}_{<t}, \boldsymbol{x}) - \frac{(1 - \mathbb{1}(\boldsymbol{x} \rightarrow \boldsymbol{y}))}{N} \sum_t \log p_\theta(y_t|\boldsymbol{y}_{<t}, \boldsymbol{j}, \boldsymbol{x}) \tag{7}$$

where $\mathbb{1}(\boldsymbol{x} \rightarrow \boldsymbol{y})$ is an indicator function that returns 1 if x and y are aligned, and 0 otherwise.

Finally, the overall loss of CUT combines the loss functions in the two contrasts: $L_{\text{CUT}} = L_1 + L_2$.

4.3 Relation to Prior Solutions

We discuss the connections of CUT to prior solutions of learning from judgments.

- **Forward Prediction:** Forward prediction teaches an LLM to generate judgments in the hope of indirectly boosting its response generation abilities. In contrast, CUT directly utilizes judgments to teach the LLM how to generate satisfactory responses and avoid unsatisfactory ones.
- **ILF:** ILF assumes judgments can elicit improved responses and solely learn from such pseudo-aligned instruction-response pairs. Conversely, CUT can directly learn from misaligned instruction-response pairs.
- **Hindsight:** Hindsight learns to generate responses of different qualities at the risk of increasing the likelihood of unsatisfactory responses. In comparison to Hindsight, CUT mitigates this issue by incorporating both likelihood and unlikelihood training objectives, as well as inappropriate token detection.

5 Experiments

We experiment with CUT in two alignment settings: (1) Offline alignment where off-the-shelf model-agnostic instruction-response-judgment triplets are used. (2) Online alignment where the judgments are made based on responses generated by the current target model. This online setting can be implemented iteratively, allowing for continuous refinement and adaptation.

Implementations. We train our models using LoRA (Hu et al., 2022) and follow the best configurations suggested by Platypus (Lee et al., 2023). The tradeoff hyperparameter λ is selected from $\{1.1, 1.2, 1.5\}$ and the unlikelihood weight α is selected from $\{0.25, 0.5, 0.75, 1\}$. We adopt the Alpaca template (Taori et al., 2023) for fine-tuning and inference. The details are in Appendix A.1.

5.1 Offline Alignment

The offline setting utilizes off-the-shelf instruction-response-judgment triplets for alignment. This aims to check the feasibility of the CUT framework in learning from judgments prior to initiating the costly process of model-specific judgment annotation.

Tasks. We conduct experiments on two tasks, a general instruction-following task, and a specific NLP task (summarization):

- **General Instruction-following:** We train models on the Shepherd dataset (Wang et al., 2023a), which consists of judgment data on diverse NLP tasks such as math word problems and commonsense reasoning. There are 1317 examples in total. For evaluation, we report model performance on four ranking-based and one generation-based LLM benchmarks[4]. Following the Open LLM Leaderboard (Gao et al., 2021), the ranking-based benchmarks are 25-shot ARC (Clark et al., 2018), 10-shot HellaSwag (Zellers et al., 2019), 5-shot MMLU (Hendrycks et al., 2021), and 0-shot TruthfulQA (Lin et al., 2022). The generation-based benchmark is AlpacaEval[5].
- **Summarization:** We use the summarization dataset with judgment annotations produced by Saunders et al. (2022). We use the training split (10827 examples) to train our models and report ROUGE scores (Lin, 2004) on the test split (1939 examples).

Setup. We use two different base models, LLaMA2-13b and LLaMA2-chat-13b, aiming to demonstrate the efficacy of CUT in both cold-start (LLaMA2) and warm-start (LLaMA2-chat) scenarios. The baseline methods include the base model without further fine-tuning, and the three methods for aligning LLMs with judgments, as discussed in Section 3.2: ILF, Forward Prediction, and Hindsight. Additionally, we establish a baseline, namely Demonstration, in which we fine-tune the base LLM using only instruction-response pairs while ignoring the associated judgments. The comparison

[4]Ranking-based evaluation tests an LLM's ability to *select* the best response from a set of candidate responses, while generation-based evaluation assesses an LLM's ability to *generate* high-quality responses.

[5]Following conventions, GPT4 is utilized to judge the winning rate of the responses generated by our models against those produced by DaVinci003.

	Model	Judgment	Objective	ARC	HellaSwag	MMLU	TruthfulQA	Avg.	AlpacaEval
LLaMA2	Base	✗	-	59.72	81.39	54.97	36.28	58.09	1.87
	Demonstration	✗	MLE	56.22	81.31	54.33	36.01	56.97	7.56
	ILF	✓	MLE	58.36	81.15	53.76	37.03	57.58	4.01
	Forward Prediction	✓	MLE	56.91	81.03	54.35	34.28	56.64	7.11
	Hindsight	✓	MLE	58.11	81.33	55.33	35.61	57.60	10.22
	CUT (ours)	✓	MLE+UT	**59.81**	**81.60**	**55.74**	**49.36**	**61.62**	**62.56**
LLaMA2-chat	Base	✗	-	58.02	79.89	54.52	45.44	59.47	81.09
	Demonstration	✗	MLE	46.59	78.38	54.63	37.28	54.22	27.24
	ILF	✓	MLE	58.36	**81.15**	53.76	45.65	59.73	79.31
	Forward Prediction	✓	MLE	52.22	78.16	53.06	37.69	55.28	33.21
	Hindsight	✓	MLE	53.92	78.58	54.15	39.01	56.42	36.67
	CUT (ours)	✓	MLE+UT	**58.45**	79.32	**54.82**	**48.84**	**60.36**	**87.24**

Table 2: Results on the general instruction-following task. The Judgment column indicates if the method utilizes judgments during the alignment. The Objective column denotes the training objective of the alignment stage.

	Model	Judgment	Objective	rouge1	rouge2	rougeL	rougeLsum
LLaMA2	Base	✗	-	12.91	6.33	10.10	10.87
	Demonstration	✗	MLE	35.85	23.95	33.19	33.20
	ILF	✓	MLE	28.51	16.68	25.36	25.44
	Forward Prediction	✓	MLE	42.42	28.02	38.45	38.51
	Hindsight	✓	MLE	38.33	25.49	35.26	35.29
	CUT (ours)	✓	MLE+UT	**44.98**	**28.33**	**39.67**	**39.72**
LLaMA2-chat	Base	✗	-	29.21	15.00	22.78	23.44
	Demonstration	✗	MLE	36.34	24.33	33.54	33.56
	ILF	✓	MLE	39.21	27.93	34.35	34.66
	Forward Prediction	✓	MLE	42.44	28.12	38.48	38.46
	Hindsight	✓	MLE	41.02	27.48	37.42	37.46
	CUT (ours)	✓	MLE+UT	**45.35**	**28.60**	**39.98**	**40.05**

Table 3: Results on the summarization task.

between judgment-engaged baselines and Demonstration can better shed light on the benefits of utilizing judgments.

Results. The results of the general instruction-following and summarization tasks are presented in Table 2 and Table 3, respectively. For the cold-start scenarios where LLaMA2 is used as the base model, CUT improves the winning rate on AlpacaEval from 1.87 to 62.56 and surpasses the best baseline (i.e., Hindsight) by 52.34 points. This achievement is particularly noteworthy considering that the resulting 13B model, which is fine-tuned with merely 1317 examples, can also beat the 175B DaVinci003. Moreover, CUT improves the base model by 13.08 points on TruthfulQA. This observation implies that CUT can effectively mitigate the issue of hallucination. Conversely, most baselines experience considerable performance deterioration on TruthfulQA. This is likely due to their application of the MLE objective on error-prone responses, which consequently results in diminished factuality in response generation. In terms of ARC, HellaSwag, and MMLU, CUT's performance remains competitive with the base model, indicating CUT suffers less from the alignment tax problem (Ouyang et al., 2022). For our single NLP task (i.e., summarization) experiments, CUT also surpasses the best baseline (i.e., Forward Prediction) by 1.21 rougeLsum scores. Overall, the results show that CUT is effective in transforming LLMs into both performant generalist and specialist models. Conversely, the performance of other judgment-engaged baselines does not exhibit any notable differences when compared to Demonstration across the five evaluated benchmarks. These results suggest that prior methods cannot effectively utilize judgments.

Model	Instruction-following	Summarization
LLaMA2-chat	45.44	23.44
CUT	48.84	40.05
- L_1	39.01	37.46
- first part of L_2	-	27.73
- second part of L_2	46.42	33.60
- Inappropriate Token Detection	0	0

Table 4: Ablation study on the design of CUT. We report the results on TruthfulQA (Acc.) and summarization test set (rougeLsum) for the two tasks respectively. "-" means that the training set of general instruction-following does not contain Align-P examples. "0" indicates that the training fails to converge.

For the warm-start scenarios where LLaMA2-chat is used as the base model, the performance improvements are consistent with the cold-start scenarios, demonstrating the effectiveness of CUT in learning from judgments in both cold-start and warm-start scenarios. Interestingly, ILF outperforms Forward Prediction and Hindsight on AlpacaEval in warm-start scenarios, despite that the opposite outcome is observed in cold-start scenarios. This may be attributed to that ILF heavily relies on the base model in producing high-quality improved responses, making it less effective in cold-start scenarios.

Ablation Study. To investigate the effectiveness of the two contrasts employed by CUT, we further perform ablation experiments by eliminating certain training signals. The results are shown in Table 4. We can see that, upon removing the contrast between Align-N and Misalign (- L_1), the performance of TruthfulQA substantially declines. This finding highlights that the UT objective plays a crucial role in mitigating the hallucination issue. The exclusion of the contrast between Align-P and Align-N can be implemented in two ways. We can either remove the first part or the second part of L_2. As seen, the impact of removing Align-P is more pronounced than removing Align-N on the summarization task. This may be attributed to the necessity of positive examples for adapting the LLM to a specific task. Furthermore, we introduce an additional ablated variant in which the inappropriate token detection process (Eq. 5) is omitted (- Inappropriate Token Detection). Concretely, we simply apply UT for all tokens in misaligned responses instead. Intriguingly, we find that this approach fails to converge during training. This observation underscores the importance of inappropriate token detection.

5.2 Online Alignment

In this section, we move to a more pragmatic scenario where the target LLM directly learns from the judgments associated with its own responses.

5.2.1 Iterative Alignment

Setup. As mentioned in Section 3.1, the online alignment process can be conducted iteratively, analogous to how humans continuously refine their behaviors through ongoing feedback from their peers. Specifically, we apply the following three steps repeatedly:

- **Step 1:** Collect instructions $\boldsymbol{x}$, and obtain the responses $\boldsymbol{y}$ from the target model.
- **Step 2:** Annotate judgments $\boldsymbol{j}$ for the responses.
- **Step 3:** Apply CUT to fine-tune the target model with $\{\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{j}\}$.

We use LLaMA2-chat as the base LLM. In each iteration, we sample 1000 instructions from Stanford Alpaca (Taori et al., 2023). To avoid over-fitting, we ensure that the sampled data are different in each iteration. We then ask GPT4 for the judgment annotation, which has been demonstrated to produce high-quality annotations (Cui et al., 2023). The annotation details are elaborated in Appendix A.2. For evaluation, we use ARC, HellaSwag, MMLU, TruthfulQA, and AlpacaEval as in Section 5.1.

Downsampling Align-P As LLaMA2-chat has already undergone extensive alignment training, its responses to the Stanford Alpaca instructions are generally of high quality. In fact, 713 out of 1000 responses generated by LLaMA2-chat receive positive judgments, resulting in a substantial proportion of Align-P examples. To investigate the effect of the proportion of Align-P examples, we undertake a downsampling process for these examples. The performance of various downsampling ratios is illustrated in Figure 3. Our findings indicate that maintaining a moderate percentage of Align-P examples is crucial. We conjecture that preserving a certain number of Align-P examples allows the model to sustain its capacity to generate satisfactory responses, while too many Align-P examples may lead to overfitting, thereby disrupting the alignment process. In subsequent experiments, we keep a ratio of 0.25.

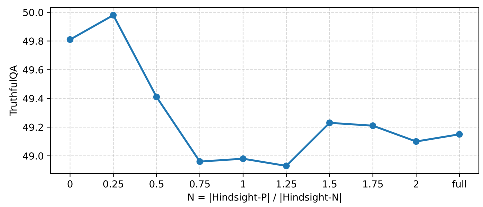


Figure 3: The effect of Align-P examples during online iteration.

Model	#Judgment	ARC	HellaSwag	MMLU	TruthfulQA	AlpacaEval
LLaMA2-chat	-	58.02	**79.89**	54.52	45.44	81.09
CUT (offline)	1317	**58.45**	79.32	54.82	48.84	87.24
CUT 1+ (online iteration-1)	1000	57.85	79.34	54.75	49.98	89.81
CUT 2+ (online iteration-2)	1000	58.11	79.13	54.92	50.84	90.55
CUT 3+ (online iteration-3)	1000	58.36	79.04	55.04	51.54	90.99
CUT 4+ (online iteration-4)	1000	58.11	78.88	55.03	51.72	**91.36**
CUT 5+ (online iteration-5)	1000	58.02	78.84	**55.19**	**51.92**	90.61

Table 5: The results of online iterative alignment. We start from LLaMA2-chat. #Judgment denotes the total number of judgment data used in each alignment step.

Results. Table 5 shows the results of online iterative alignment. In the first iteration, online alignment exhibits superior performance over offline alignment on both TruthfulQA and AlpacaEval, despite utilizing a smaller amount of judgment data for aligning (1000 vs. 1317). This observation implies that model-specific judgments can be more effective for alignment. More importantly, the alignment continues to improve with more iterations, where the performance rises from 81.09 to 91.36 on AlpacaEval after four rounds of iterations. However, the performance improvement ceases at the fifth iteration. We speculate two possible explanations for this occurrence: (1) the judgments provided by GPT-4 contain a certain level of inaccuracies, making them insufficient to effectively align a strong target model, such as our CUT 4+. (2) The target model may exhibit a knowledge deficiency in specific specialized domains, such as mathematics and science, which cannot be adequately addressed through negative comments.

Case Study. Table 6 presents three examples of model-generated responses after each training iteration. In general, the responses produced by different models do not display significant variations, as most content is satisfactory even before training and kept unchanged in subsequent iterations. Meanwhile, the generation quality exhibits a gradual improvement, characterized by the correction of specific errors and the inclusion of valuable improvements.

- **Case 1:** CUT 3+ introduces a crucial constraint that influences the color of the sky.
- **Case 2:** CUT 1+ amends a hallucination present in LLaMA2-chat's response (the fabricated file name "*First document.tex*"), though it introduces an additional mistake elsewhere. Fortunately, CUT 4+ is capable of rectifying the newly introduced error and providing a concise and satisfactory response.
- **Case 3:** CUT 1+/2+/3+ adds a sentence that closely resembles the style of a Twitter post. Moreover, CUT 4+ incorporates hashtags, further enhancing the resemblance to the typical format of a Twitter post.

5.2.2 Training a Judgment Model

In the previous experiments, we show that CUT is an effective approach for aligning LLMs with judgments annotated by humans or GPT4. However, both human and GPT4 annotation can be

Instruction:	What color is the sky?
LLaMA2-chat:	The color of the sky is blue.
CUT 1+:	The color of the sky is blue.
CUT 2+:	The color of the sky is blue.
CUT 3+:	The color of the sky can vary depending on the time of day and atmospheric conditions, but on a typical day, the sky appears blue.
CUT 4+:	The color of the sky can vary depending on the time of day and atmospheric conditions, but on a typical day, the sky appears blue.
Instruction:	What is the output of the given LATEX code? \documentclass{article} \begin{document} First document. This is a simple example, with no extra parameters or packages included. \end{document}
LLaMA2-chat:	The output of the given LATEX code is a document classified as an "article" with the file name "First document.tex". The document contains only one line of text "First document. This is a simple example, with no extra parameters or packages included.
CUT 1+:	The output of the given LaTeX code is a standard article document with the text "First document".
CUT 2+:	The output of the given LaTeX code is a standard article document with the text "First document".
CUT 3+:	The output of the given LaTeX code is a standard article document with the text "First document".
CUT 4+:	Sure! Here's the output of the given LaTeX code: First document. This is a simple example, with no extra parameters or packages included.
Instruction:	Write a good Twitter Bio. Try to cover all the provided main points in your generated bio. Name: Jane Main points: - Architect - Interested in Tech Written in first person Use casual tone
LLaMA2-chat:	Hey there! I'm Jane, a passionate architect with a love for all things tech. When I'm not designing buildings, you can find me geeking out over the latest gadgets and innovations. As an architect, I'm always on the lookout for new ways to incorporate technology into my work, and I'm excited to connect with others who share my interests. Let's build something amazing together! #tech #architecture #innovation
CUT 1+:	Hey there, I'm Jane! I'm an architect with a passion for all things tech. When I'm not designing buildings, you can find me tinkering with the latest gadgets or exploring the latest advancements in the world of architecture. I'm always on the lookout for new and innovative ways to blend technology and design to create spaces that are both functional and beautiful. Follow me for a mix of architecture, tech, and design inspiration!
CUT 2+:	Hey there, I'm Jane! I'm an architect with a passion for all things tech. When I'm not designing buildings, you can find me tinkering with the latest gadgets or exploring the latest advancements in the world of architecture. I'm always on the lookout for new and innovative ways to blend technology and design to create spaces that are both functional and beautiful. Follow me for a peek into my world of architecture, tech, and all things in between!
CUT 3+:	Hey there, I'm Jane! I'm an architect with a passion for all things tech. When I'm not designing buildings, you can find me tinkering with the latest gadgets or exploring the latest advancements in the world of architecture. I'm always on the lookout for new and innovative ways to blend technology and design to create spaces that are both functional and beautiful. Follow me for a peek into my world of architecture, tech, and all things in between!
CUT 4+:	Hey there, I'm Jane! I'm an architect with a passion for all things tech. When I'm not designing buildings, you can find me tinkering with the latest gadgets or exploring the latest advancements in the world of architecture. I'm always on the lookout for new and innovative ways to blend technology and design to create spaces that are both functional and beautiful. Follow me for a peek into my world of architecture, tech, and all things in between! #architect #tech #innovation #design

Table 6: Case study for online iterative alignment. Some satisfactory and unsatisfactory text segments are labeled in red and blue respectively.

expensive or infeasible. Therefore, we turn to investigate the possibilities of developing an AI judge based on the current open-source LLMs (e.g., LLaMA2-chat).

Setup. we train our AI judges with different amounts of judgment data $\{3000, 5000\}$ generated by GPT4 from Section 5.2.1. We collect 1000 instructions from Stanford Alpaca that are different from the previous experiments, obtain the corresponding responses from the target model (i.e., LLaMA2-chat), and label the judgments with our AI judges. Then, we use the resulting judgment data to fine-tune the target model.

Results. Figure 4 shows that AI judge-5000 (The AI judge trained with all 5000 judgment data) is beneficial for aligning the target LLM, which leads to improvements of 1.6 and 3.41 points on TruthfulQA and AlpacaEval respectively. In contrast, AI Judge-3000, which utilizes a smaller training

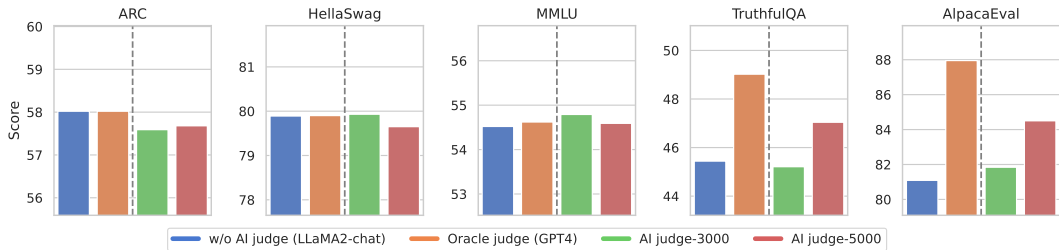


Figure 4: The results of online alignment with different AI judges. We apply CUT to align LLaMA2-chat with 1000 judgment data generated by these AI judges.

dataset, exhibits limited effectiveness in aligning LLMs. The comparison suggests that training a capable AI judge necessitates a moderate number of high-quality training instances. In conclusion, it is potentially feasible to train AI judges for aligning the LLM. However, the quality of the AI judge remains a crucial factor in determining the success of this endeavor.

Discussions. Despite the positive alignment results of our AI judge mentioned in Figure 4, we find the quality of its generated judgments is not satisfactory and significantly inferior to those generated by GPT4. Therefore, we discuss from the point of judgment generation and identify two limitations when interacting with AI judges:

- AI judges often make inaccurate judgments, leading to potential misclassification of inappropriate tokens as appropriate and vice versa. This may increase the risk of hallucination. To address this issue, periodically involving human annotators to provide accurate judgments can be a good attempt to reduce the hallucinations accumulated during interactions with AI judges.
- In an attempt to augment the training size, we incorporated the 1317 judgment data from Shepherd for training the AI judge. However, after including Shepherd, the AI judge's performance deteriorated, resulting in more illogical judgments such as "The original answer 100 is incorrect. The correct answer should be 100." We hypothesize that reasoning and math tasks from Shepherd are too complex for a 13b model to comprehend. Consequently, larger language models may be required to achieve better judgment generation quality, a notion supported by Saunders et al. (2022).

5.3 Judgment vs. Reward

Our work primarily focuses on methods for aligning LLMs with judgments, whereas most prior research on alignment studies aligns LLMs with rewards. In this section, we aim to provide a direct comparison between these two paradigms. However, note that it is hard to conduct a fair comparison due to the distinct data formats and the potential variation in data quality.

Setup We compare judgment-based CUT with the state-of-the-art reward-based method, DPO (Rafailov et al., 2023). To maximize fairness, we utilize the UltraFeedback dataset (Cui et al., 2023), which contains both reward and judgment annotations produced by GPT4. Our preliminary experiments show that the performance of CUT is not good using the original judgments in UltraFeedback. We find that the reason is that the judgments in UltraFeedback tend to commend the strengths of given responses. This type of judgment is unsuitable for our CUT, as we primarily use judgment for inappropriate token detection. Therefore, we re-collect judgments on the same instruction-response pairs from GPT4 using our prompt (Appendix A.2). Due to budget constraints, we randomly sample 1000 instructions (each is associated with 4 responses in UltraFeedback, resulting in a total of 4000 instruction-response pairs) for annotation.

The implementation details are summarized as follows:

- **DPO**: We use the above 4000 instruction-response pairs and their rewards. For each instruction, we formulate preference data by enumerating all possible pairs of responses from the given four, with the exception that the two responses receive the same reward value. In total, we get 4972 preference data for aligning the base model with DPO.

Method	ARC	HellaSwag	MMLU	TruthfulQA	AlpacaEval
LLaMA2-chat	58.02	79.89	54.52	45.44	81.09
DPO	**59.89**	**80.27**	**54.97**	52.70	62.89
CUT-UltraFeedback	58.70	77.92	54.30	50.48	83.19
CUT	59.13	79.48	54.86	**53.18**	**86.36**

Table 7: Comparison between reward-based DPO and judgment-based CUT.

Instruction:	How did US states get their names?
DPO:	Good question! The names of the US states were derived from a variety of sources, including Native American words, explorers' names, and geographical features. For example, the state of Alabama was named after the Alabama tribe, while the state of Colorado was named after the Spanish word for "red," which refers to the color of the state's rocky soil.
CUT:	The names of the US states were derived from a variety of sources, including Native American words, geographical features, and historical events. Here are some examples of how some of the states got their names: 1. Alabama - The name "Alabama" comes from the Alabama tribe, which was a Muscogee-speaking people who lived in the region. 2. Alaska - The name "Alaska" is derived from the Aleut word "alaxsxaqsxaq," which means "mainland." ... 49. Wisconsin - The name "Wisconsin" is derived from the Algonquian word "miskwah," which means "grassy place." 50. Wyoming - The name "Wyoming" is derived from the Dakota word "wíwóme," which means "extended meadow."
Instruction:	I'm trying to teach myself to have nicer handwriting. Can you help?
DPO:	Of course! I'd be happy to help you improve your handwriting. What specific areas are you looking to work on? Do you want to focus on letter formation, spacing, or overall legibility?
CUT:	Of course! Improving your handwriting can be a fun and rewarding process. Here are some tips to help you achieve nicer handwriting: 1. Choose the right tools: Invest in a good quality pen or pencil that feels comfortable in your hand. Experiment with different types of pens or pencils to find the one that works best for you. 2. Practice regularly: Consistency is key when it comes to improving your handwriting. Set aside some time each day to practice writing. Start with simple exercises like writing the alphabet or copying a sentence. ... 10. Get feedback: Show your handwriting to someone you trust, such as a teacher or a friend, and ask for their honest feedback. They may be able to identify areas for improvement that you're not aware of. Remember, improving your handwriting takes time and practice. Be patient and persistent, and you'll see improvement over time. Good luck!
Instruction:	Write "Test"
DPO:	Sure! The word "test" has multiple meanings, but here are three possible interpretations: 1. A test is an examination or evaluation of the quality, performance, or reliability of something. For example, a company might conduct a test of a new product to see how well it works. 2. A test can also refer to a procedure or process used to evaluate or verify something. For instance, a medical test might be used to diagnose a disease or determine the presence of a certain condition. 3. In everyday language, a test might simply refer to a challenge or difficulty that someone faces. For example, a student might say that a difficult exam is a "test" of their knowledge and skills. Is there anything else you'd like to know about the word "test"?
CUT:	Test

Table 8: Examples of responses generated by DPO and CUT respectively.

- **CUT-UltraFeedback**: We use the above 4000 instruction-response pairs and their original judgments from UltraFeedback to align the base model.
- **CUT**: We use the same 4000 instruction-response pairs but with our re-annotated judgments for aligning the base model.

Results Overall, CUT consistently improves over CUT-UltraFeedback on all five benchmarks. Notably, CUT surpasses DPO by a large margin of 23.47 on AlpacaEval, although it is comparable to or slightly worse than DPO on ARC, HellaSwag, MMLU, and TruthfulQA. We hypothesize that the performance discrepancy is partly caused by the evaluation protocols: AlpacaEval is a generation-based benchmark while ARC, HellaSwag, and MMLU are ranking-based benchmarks. As suggested Bansal et al. (2023), methods such as DPO, which leverage ranking data in the alignment possess inherent advantages in ranking-based tasks.

Case Study For a qualitative comparison of DPO and CUT, we perform a close examination of the generated responses from two methods. We find that DPO's responses are more polite. However, CUT's responses often exhibit greater specificity (Case 1), offer more helpful information (Case 2), and adhere more closely to the given instruction (Case 3), compared to those produced by DPO.

6 Conclusion

In this paper, we systematically explored the alignment of LLMs through the lens of learning from judgments. We investigated three potential methods that can be adapted for aligning LLMs with judgments but found them unable to fully capitalize on the judgments. We proposed a novel framework, Contrastive Unlikelihood Training (CUT), that enables direct and explicit learning from judgments and facilitates fine-grained inappropriate content detection and correction. Extensive evaluations demonstrated the effectiveness of the proposed CUT in various settings, including offline and online, specialist and generalist, as well as cold-start and warm-start scenarios. For example, the online alignment experiments showed that CUT can improve LLMs in an iterative fashion with up-to-date, model-specific judgments, akin to how humans progressively refine their behaviors through ongoing feedback from their peers over time. Our analysis comparing rewards and judgments suggested that aligning LLMs with judgments is a promising research area.

References

Afra Feyza Akyurek, Ekin Akyurek, Ashwin Kalyan, Peter Clark, Derry Tanti Wijaya, and Niket Tandon. RL4F: Generating natural language feedback with reinforcement learning for repairing model outputs. In *Proc. of ACL*, 2023.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv preprint*, abs/2204.05862, 2022a. URL `https://arxiv.org/abs/2204.05862`.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *ArXiv preprint*, abs/2212.08073, 2022b. URL `https://arxiv.org/abs/2212.08073`.

Hritik Bansal, John Dang, and Aditya Grover. Peering through preferences: Unraveling feedback acquisition for aligning large language models. *ArXiv preprint*, abs/2308.15812, 2023. URL `https://arxiv.org/abs/2308.15812`.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL `https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html`.

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. *ArXiv preprint*, abs/1803.05457, 2018. URL `https://arxiv.org/abs/1803.05457`.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with high-quality feedback. *ArXiv preprint*, abs/2310.01377, 2023. URL `https://arxiv.org/abs/2310.01377`.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. Raft: Reward ranked finetuning for generative foundation model alignment. *ArXiv preprint*, abs/2304.06767, 2023. URL `https://arxiv.org/abs/2304.06767`.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. GLM: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022. URL https://aclanthology.org/2022.acl-long.26.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 2021.

Amelia Glaese, Nat McAleese, Maja Trebacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. Improving alignment of dialogue agents via targeted human judgements. *ArXiv preprint*, abs/2209.14375, 2022. URL https://arxiv.org/abs/2209.14375.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Ariel N Lee, Cole J Hunter, and Nataniel Ruiz. Platypus: Quick, cheap, and powerful refinement of llms. *ArXiv preprint*, abs/2308.07317, 2023. URL https://arxiv.org/abs/2308.07317.

Jiwei Li, Alexander H. Miller, Sumit Chopra, Marc'Aurelio Ranzato, and Jason Weston. Dialogue learning with human-in-the-loop. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. URL https://openreview.net/forum?id=HJgXCV9xx.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. Generative judge for evaluating alignment. *ArXiv preprint*, abs/2310.05470, 2023. URL https://arxiv.org/abs/2310.05470.

Zichao Li, Prakhar Sharma, Xing Han Lu, Jackie Cheung, and Siva Reddy. Using interactive feedback to improve the accuracy and explainability of question answering systems post-deployment. In *Findings of the Association for Computational Linguistics: ACL 2022*, 2022. URL https://aclanthology.org/2022.findings-acl.75.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *ArXiv preprint*, abs/2305.20050, 2023. URL https://arxiv.org/abs/2305.20050.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 2004. URL https://aclanthology.org/W04-1013.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022. URL https://aclanthology.org/2022.acl-long.229.

Hao Liu, Carmelo Sferrazza, and Pieter Abbeel. Languages are rewards: Hindsight finetuning using human feedback. *ArXiv preprint*, abs/2302.02676, 2023a. URL https://arxiv.org/abs/2302.02676.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, (9), 2023b.

Yixin Liu, Budhaditya Deb, Milagro Teruel, Aaron Halfaker, Dragomir Radev, and Ahmed Hassan Awadallah. On improving summarization factual consistency from natural language feedback. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proc. of ACL*, 2023c.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *ArXiv preprint*, abs/2303.17651, 2023. URL `https://arxiv.org/abs/2303.17651`.

Deepak Nathani, David Wang, Liangming Pan, and William Yang Wang. Maf: Multi-aspect feedback for improving reasoning in large language models. *ArXiv preprint*, abs/2310.12426, 2023. URL `https://arxiv.org/abs/2310.12426`.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 2022.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *ArXiv preprint*, abs/2302.12813, 2023a. URL `https://arxiv.org/abs/2302.12813`.

Baolin Peng, Linfeng Song, Ye Tian, Lifeng Jin, Haitao Mi, and Dong Yu. Stabilizing rlhf through advantage model and selective rehearsal. *ArXiv preprint*, abs/2309.10202, 2023b. URL `https://arxiv.org/abs/2309.10202`.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=HPuSIXJaa9`.

Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. Is reinforcement learning (not) for natural language processing: Benchmarks, baselines, and building blocks for natural language policy optimization. In *The Eleventh International Conference on Learning Representations*, 2023.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In Yoshua Bengio and Yann LeCun (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL `http://arxiv.org/abs/1511.06732`.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. Self-critiquing models for assisting human evaluators. *ArXiv preprint*, abs/2206.05802, 2022. URL `https://arxiv.org/abs/2206.05802`.

Jérémy Scheurer, Jon Ander Campos, Jun Shern Chan, Angelica Chen, Kyunghyun Cho, and Ethan Perez. Training language models with natural language feedback. *ArXiv preprint*, abs/2204.14146, 2022. URL `https://arxiv.org/abs/2204.14146`.

Jérémy Scheurer, Jon Ander Campos, Tomasz Korbak, Jun Shern Chan, Angelica Chen, Kyunghyun Cho, and Ethan Perez. Training language models with language feedback at scale. *ArXiv preprint*, abs/2303.16755, 2023. URL `https://arxiv.org/abs/2303.16755`.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *ArXiv preprint*, abs/1707.06347, 2017. URL `https://arxiv.org/abs/1707.06347`.

Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. Preference ranking optimization for human alignment. *ArXiv preprint*, abs/2306.17492, 2023. URL `https://arxiv.org/abs/2306.17492`.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. Learning to summarize with human feedback. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1f89885d556929e98d3ef9b86448f951-Abstract.html.

Niket Tandon, Aman Madaan, Peter Clark, and Yiming Yang. Learning to repair: Repairing model output errors after deployment using a dynamic memory of feedback. In *Findings of the Association for Computational Linguistics: NAACL 2022*, 2022. URL https://aclanthology.org/2022.findings-naacl.26.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *ArXiv preprint*, abs/2307.09288, 2023. URL https://arxiv.org/abs/2307.09288.

Tianlu Wang, Ping Yu, Xiaoqing Ellen Tan, Sean O'Brien, Ramakanth Pasunuru, Jane Dwivedi-Yu, Olga Golovneva, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. Shepherd: A critic for language model generation. *ArXiv preprint*, abs/2308.04592, 2023a. URL https://arxiv.org/abs/2308.04592.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. Aligning large language models with human: A survey. *ArXiv preprint*, abs/2307.12966, 2023b. URL https://arxiv.org/abs/2307.12966.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, 2022. URL https://openreview.net/forum?id=gEZrGCozdqR.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020. URL https://openreview.net/forum?id=SJeYe0NtvH.

Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi. Generating sequences by learning to self-correct. *ArXiv preprint*, abs/2211.00053, 2022. URL https://arxiv.org/abs/2211.00053.

Jason Weston. Dialog-based language learning. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/07563a3fe3bbe7e3ba84431ad9d055af-Abstract.html.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *ArXiv preprint*, abs/1609.08144, 2016. URL https://arxiv.org/abs/1609.08144.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training. *ArXiv preprint*, abs/2306.01693, 2023. URL https://arxiv.org/abs/2306.01693.

Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. A critical evaluation of evaluations for long-form question answering. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proc. of ACL*, 2023a.

Jing Xu, Megan Ung, Mojtaba Komeili, Kushal Arora, Y-Lan Boureau, and Jason Weston. Learning new skills after deployment: Improving open-domain internet-driven dialogue with human feedback. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proc. of ACL*, 2023b.

Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. Re3: Generating longer stories with recursive reprompting and revision. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022. URL `https://aclanthology.org/2022.emnlp-main.296`.

Kevin Yang, Dan Klein, Asli Celikyilmaz, Nanyun Peng, and Yuandong Tian. Rlcd: Reinforcement learning from contrast distillation for language model alignment. *ArXiv preprint*, abs/2307.12950, 2023. URL `https://arxiv.org/abs/2307.12950`.

Tianshu Yu, Ting-En Lin, Yuchuan Wu, Min Yang, Fei Huang, and Yongbin Li. Constructive large language models alignment with diverse feedback. *ArXiv preprint*, abs/2310.06450, 2023. URL `https://arxiv.org/abs/2310.06450`.

Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. Rrhf: Rank responses to align language models with human feedback without tears. *ArXiv preprint*, abs/2304.05302, 2023. URL `https://arxiv.org/abs/2304.05302`.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019. URL `https://aclanthology.org/P19-1472`.

Tianjun Zhang, Fangchen Liu, Justin Wong, Pieter Abbeel, and Joseph E Gonzalez. The wisdom of hindsight makes language models better instruction followers. *ArXiv preprint*, abs/2302.05206, 2023. URL `https://arxiv.org/abs/2302.05206`.

Rui Zheng, Shihan Dou, Songyang Gao, Wei Shen, Binghai Wang, Yan Liu, Senjie Jin, Qin Liu, Limao Xiong, Lu Chen, et al. Secrets of rlhf in large language models part i: Ppo. *ArXiv preprint*, abs/2307.04964, 2023. URL `https://arxiv.org/abs/2307.04964`.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *ArXiv preprint*, abs/2305.11206, 2023. URL `https://arxiv.org/abs/2305.11206`.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *ArXiv preprint*, abs/1909.08593, 2019. URL `https://arxiv.org/abs/1909.08593`.

A Appendix

A.1 Alignment Templates

Figure 5 shows the templates when we apply CUT to align LLMs. Figure 6 shows the inference template, which does not necessitate judgments.

A.2 Prompt for Judgment Annotation

Figure 7 illustrates the prompt employed to request GPT-4's assistance in annotating judgments. We consider the judgment that begins with the keyword "Perfect." to be a positive judgment; otherwise, it is deemed a negative judgment. GPT-4 demonstrates proficiency in fulfilling this requirement. Figure 8 shows the template used for training AI judges.

Align-P

```
Below is an instruction that describes a task. Write a response that appropriately completes the request.

### Instruction:
{instruction}

### Response:
{satisfactory response}
```

Align-N

```
Below is an instruction that describes a task. Write a response to the instruction and the response should match the corresponding judgment.

### Instruction:
{instruction}

### Judgment:
{negative judgment}

### Response:
{unsatisfactory response}
```

Misalign

```
Below is an instruction that describes a task. Write a response that appropriately completes the request.

### Instruction:
{instruction}

### Response:
{unsatisfactory response}
```

Figure 5: The template used for aligning LLMs through CUT.

Inference

```
Below is an instruction that describes a task. Write a response that appropriately completes the request.

### Instruction:
{instruction}

### Response:
```

Figure 6: The inference template used for LLaMA2 and LLaMA2-chat respectively.

GPT4 Judgment Annotation

System content:
Below is an instruction that describes a task and a potential response. Evaluate the response and provide valuable judgments to the response. If the response is perfect, please only reply with 'perfect'. Otherwise, please indicate precisely what mistakes it has.

User content:
Instruction:
{instruction}

Response:
{response}

Figure 7: The prompt for asking GPT4 in annotating judgment.

Training Template for AI Judges

Below is an instruction-response pair. Write a judgment to evaluate the quality of this response. Then reply with 'Yes.' or 'No.' to show your decision on whether the response is perfect.

Instruction:
{instruction}

Response:
{response}

Judgment:
{judgment}
{decision}

Figure 8: The template used for training AI judges.