Diverse AI Feedback For Large Language Model Alignment

Tianshu Yu[123*], **Ting-En Lin**[3], **Yuchuan Wu**[3], **Min Yang**[14†], **Fei Huang**[3], **Yongbin Li**[3†]
[1]Shenzhen Key Laboratory for High Performance Data Mining,
Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, China
[2]University of Chinese Academy of Sciences, China [3]Tongyi Lab, China
[4]Shenzhen University of Advanced Technology, China
{ts.yu, min.yang}@siat.ac.cn
{ting-en.lte, shuide.lyb}@alibaba-inc.com

Abstract

Recent advances in large language models (LLMs) focus on aligning models with human values to minimize harmful content. However, existing methods often rely on a single type of feedback, such as preferences, annotated labels, or critiques, which can lead to overfitting and suboptimal performance. In this paper, we propose **D**iverse **AI** **F**eedback (DAIF), a novel approach that integrates three types of feedback—critique, refinement, and preference—tailored to tasks of varying uncertainty levels. Through an analysis of information gain, we show that critique feedback is most effective for low-uncertainty tasks, refinement feedback for medium-uncertainty tasks, and preference feedback for high-uncertainty tasks. Training with this diversified feedback reduces overfitting and improves alignment. Experimental results across three tasks—question answering, dialog generation, and text summarization–demonstrate that DAIF outperforms traditional methods relying on a single feedback type.[1]

1 Introduction

In recent years, large language models (LLMs) have demonstrated significant capabilities in addressing a wide range of information needs (Bubeck et al., 2023; Touvron et al., 2023; Li et al., 2023; Muennighoff et al., 2023; Tao et al., 2024; Luo et al., 2024). A growing area of research in this field focuses on *aligning LLMs with human values* to reduce the risk of generating harmful or misleading content (Bai et al., 2022a; Wang et al., 2023b). This objective has been the focus of numerous studies (Stiennon et al., 2020; Ouyang et al., 2022; Akyürek et al., 2023; Wu et al., 2023). To achieve this, most current alignment methods rely on human-annotated (Ziegler et al., 2019; Nakano et al., 2021) or AI-generated (Bai et al., 2022a; Chen et al., 2024) preference feedback, which is used in supervised training (Stiennon et al., 2020; Rafailov et al., 2023) or reinforcement learning (Ouyang et al., 2022).

However, a key limitation of these existing alignment methods is their tendency to overfit (Gao et al., 2023). To address this issue, recent research has explored alternative forms of feedback, such as refinement feedback, which directly refines model outputs (Shi et al., 2022; Welleck et al., 2022), and critique feedback, which provides natural language critiques to guide the model in self-improvement (Tandon et al., 2021; Scheurer et al., 2022; Madaan et al., 2023). Despite these efforts, these methods have not fully resolved the overfitting problem. Previous studies suggest that the overfitting of LLM alignment is largely due to the reliance on a single feedback type, which leads to low data efficiency.

Previous research in the field of robotics has laid a strong foundation for the integration of multiple feedback mechanisms. For example, Jeon et al. (2020) examined the feasibility of combining various feedback types for reward learning, highlighting the potential advantages of such an approach. Similarly, Bıyık et al. (2022) successfully utilized human demonstrations and preference feedback to enhance reward function learning, further demonstrating the value of incorporating diverse feedback sources. Building on this, Ghosal et al. (2023) explored the optimization of feedback selection, illustrating the effectiveness of

[*] This work was conducted when Tianshu Yu was interning at Tongyi Lab.

[†] Min Yang and Yongbin Li are corresponding authors.

[1]Our code and data are available at: https://github.com/AlibabaResearch/DAMO-ConvAI/tree/main/DAIF.

Transactions of the Association for Computational Linguistics, vol. 13, pp. 392–407, 2025. https://doi.org/10.1162/tacl_a_00746
Action Editor: Alexander Rush. Submission batch: 12/2023; Revision batch: 7/2024; Published 4/2025.

combining different types of feedback to improve agent policies. However, a notable gap in the literature exists regarding the application of these findings to the alignment of LLMs. Moreover, existing studies have yet to investigate the most suitable feedback type for each training sample, which could maximize training data efficiency.

In response to these gaps, we introduce Diverse AI Feedback (DAIF), a novel approach that optimizes LLM alignment by integrating three types of feedback: critique (**Critic**), refinement (**Refine**), and preference (**Prefer**). DAIF tailors these feedback types to tasks of varying uncertainty to enhance alignment with human values. Starting with an unaligned base model, we generate outputs for a predefined set of problems. Using the concept of information gain from active learning (Bıyık et al., 2019), we evaluate the information gain for each feedback type across different uncertainty levels of problems, with perplexity, a property that has been widely recognized in studies addressing uncertainty in language generation tasks (Baan et al., 2023; Hu et al., 2023), serves as a task-agnostic metric for assessing uncertainty. Based on this analysis, we classify problems into three uncertainty categories: ''low'', ''medium'', and ''high''. For ''low'' tasks, critique feedback, consisting of natural language critiques, is utilized; for ''medium'' tasks, refinement feedback, involving improvements to the model outputs, is employed; and for the most challenging ''high'' tasks, preference feedback, drawn from annotated preferences across multiple outputs generated by the same model, is sought.

The primary contributions of our work are as follows:

1. We introduce DAIF as a novel method for enhancing the alignment process by integrating diverse feedback mechanisms.

2. We investigate the optimal strategy for combining critique, refinement, and preference feedback types, using an information gain approach to determine the best fit for the training dataset.

3. We propose a differentiated feedback approach that tailors the use of various feedback types to the uncertainty levels of the tasks at hand.

4. We present experimental results across three downstream tasks, demonstrating that DAIF outperforms traditional methods relying on a single feedback type. Additional experiments and analyses further validate the effectiveness of DAIF.

2 Related Work

2.1 Research on the Alignment of LLMs

In recent years, the task of fine-tuning language models to align with human values has gained paramount importance, driven by the imperative to reduce the generation of incorrect, misleading, or harmful content in dialog completions (Bai et al., 2022a; Liu et al., 2023b; Wang et al., 2023b; Gao et al., 2024). Reinforcement learning (RL) has become the predominant technique in numerous prior studies tackling this challenge. RL frames the generation process as a Markov decision process and optimizes the policy model to maximize a proxy reward, establishing itself as a pivotal method in this context. For instance, Ziegler et al. (2019) were pioneers in investigating the RLHF method for stylistic continuation and summary generation. Bai et al. (2022b) introduced the concept of LLM alignment along with the HHH (helpful, harmless, honest) principle, applying RLHF to achieve alignment. Ouyang et al. (2022) introduced InstructGPT, which was subsequently applied to the renowned ChatGPT. In addition to RLHF, alternative training methods have been explored. Liu et al. (2023a) proposed CoH, which learned from both good and bad responses. Rafailov et al. (2023) introduced the DPO algorithm to mitigate the instability of PPO training, derived from the classic Bradley-Terry model of reward proxy learning. Song et al. (2023) extended DPO to scenarios where a prompt can elicit more than two possible responses with annotated human preference order.

2.2 Learning from Various Types of Feedback

Existing literature has explored diverse forms of feedback to enhance model predictive capabilities. These methods can be classified as (i) preferences that involve pairwise comparisons or rankings (Bai et al., 2022a; Gao et al., 2022; Zhu et al., 2023; Feng et al., 2024); (ii) natural language critiques (Tandon et al., 2021; Scheurer

et al., 2022; Saunders et al., 2022; Madaan et al., 2023); and (iii) direct textual refinements of generated outputs (Shi et al., 2022; Welleck et al., 2022). Saunders et al. (2022) introduced a learning paradigm known as Self-Critique, where a model evaluated its own outputs in natural language and then refined itself based on these critiques, while Chen et al. (2024) and Wu et al. (2024) studied Self-Play, which encourages a model to improve its ability by playing against instances of itself. Ethayarajh et al. (2024) and Jung et al. (2024) proposed algorithms to align LLMs by simple ''accept'' and ''disapproving'' signals. Specialized feedback types have also been developed for task-specific applications. For instance, Gao et al. (2022) employed accuracy metrics in extractive question-answering as feedback for policy fine-tuning, while Uesato et al. (2022) used the correctness of both the solution process and the final outcomes as feedback.

2.3 Active Learning for LLM Training

Active learning (Settles, 2009) has gained significant attention in improving the data efficiency of training LLMs, particularly in scenarios where labeled data is scarce or expensive. Gleave and Irving (2022) attempted to apply active learning on language reward modeling using ensemble-based methods and Thompson sampling to reduce the amount of training data. Mehta et al. (2023) formalized the problem of RLHF into dueling contextual bandit learning and developed an active exploration method that samples the most informative prompt and answer pair for preference labeling. Das et al. (2024) derived a method for active problem selection and answer pair sampling under the framework of Bradley-Terry model (Bradley and Terry, 1952). Melo et al. (2024) improved the traditional Bayesian active learning algorithm for preference modeling, considering both feature space entropy and preference model uncertainty. However, those methods mainly focuses on training with preference feedback, and most of them require additional training to determine the uncertainty of problems.

Unlike prior methods, we present DAIF to align LLMs with human values through a differentiated feedback mechanism. DAIF distinguishes itself by incorporating three distinct types of feedback, each calibrated to the problem's uncertainty level, enhancing the effectiveness of model training.

3 Preliminary

In this section, we first introduce the dataset used in our experiments, along with the various feedback types employed. We then present the information gain methodology utilized to determine the optimal feedback type for each problem. Subsequently, we conduct a preliminary study to evaluate the information gain associated with critique, refinement, and preference feedback types. Based on the results of this analysis, we propose an effective strategy for combining these diverse feedback mechanisms.

3.1 The Construction of Problem Dataset

To commence our study, we assemble a problem dataset, denoted as $\mathcal{D}$, for the purposes of feedback collection and model training. We consider three specific datasets for our three distinct tasks within the realm of natural language processing: the *WebGPT-comparison* dataset for question answering, the *HH-RLHF* dataset for dialogue generation, and the *OpenAI-Summarize-TLDR* dataset for text summarization. We randomly select 10,000 samples from each of these datasets and combine them to create a comprehensive problem dataset $\mathcal{D}$, which forms the experimental foundation of our study.

3.2 Feedback Types Description

Initially, in our experiment, we use the *Vicuna-7B* model as our base model. For a given problem q in our dataset $\mathcal{D}$, we first get an initial answer a by greedy sampling from our base model. Then, we collect each type of feedback following the methods given below:

- Critique: This feedback type provides constructive suggestions for enhancing answer a, enabling our base model to further refine itself. We solicit the API to provide improvement suggestions s for answer a. Subsequently, the base model generates an improved answer a_c based on q, a, and s.
- Refinement: In this feedback category, we receive an improved version of the a. We make an API query to enhance the answer a, resulting in the refined answer a_r.
- Preference: In this approach, the model selects the superior answer among two distinct answers generated for the same prompt. To

this end, we follow the traditional process of RLHF to sample two distinct answers a_1, a_2 under question q from our base model. Subsequently, we employ the API to determine whether a_1 or a_2 is the better option.

3.3 The Information Gain Framework

The information gain approach has demonstrated its effectiveness and superiority over traditional active learning methods (Palan et al., 2019), by selecting the comparison pair (a^+, a^-) that maximizes the information gain $IG(a^+, a^-|q)$ as defined by the following equation:

$$IG(a^+, a^-|q) = H(a^+, a^-|q) - H(a^+, a^-|q, \pi^*)$$

Here, $H(a^+, a^-|q)$ represents the entropy associated with the base model π determining that answer a^+ is preferred over a^-, while $H(a^+, a^-|q, \pi^*)$ denotes the entropy under the optimal model π^* where a^+ is preferred over a^-:

$$\begin{aligned} H(a^+, a^-|q) = &- p(a^+ \succ a^-; \pi) \log p(a^+ \succ a^-; \pi) \\ &- p(a^- \succ a^+; \pi) \log p(a^- \succ a^+; \pi) \\ H(a^+, a^-|q, \pi^*) = &- p(a^+ \succ a^-; \pi^*) \log p(a^+ \succ a^-; \pi^*) \\ &- p(a^- \succ a^+; \pi^*) \log p(a^- \succ a^+; \pi^*) \end{aligned}$$

By combining the Bradley-Terry model (Bradley and Terry, 1952) with the energy-based model (LeCun et al., 2006), we derive the following expressions for $p(a^+ \succ a^-; \pi)$ and $p(a^+ \succ a^-; \pi^*)$:

$$\begin{aligned} p(a^+ \succ a^-; \pi) &= \frac{\pi(a^+|q)}{\pi(a^+|q) + \pi(a^-|q)} \\ p(a^+ \succ a^-; \pi^*) &= \frac{\pi^*(a^+|q)}{\pi^*(a^+|q) + \pi^*(a^-|q)} \end{aligned}$$

We adopt the optimal solution from Rafailov et al. (2023), where $\pi^*(a|q) \propto \pi(a|q) \exp\left(\frac{1}{\beta} R(a|q)\right)$, as the representation of the optimal policy π^*. Hence, we have

$$\begin{aligned} \pi^*(a^+|q) &= \pi(a^+|q) \exp\left(\frac{1}{\beta} R(a^+|q)\right) \\ \pi^*(a^-|q) &= \pi(a^-|q) \exp\left(\frac{1}{\beta} R(a^-|q)\right) \end{aligned}$$

In our experiment, all three types of feedback are converted into comparison pairs. This allows us to compute the information gain for each feedback type and subsequently select the feedback type that maximizes the information gain.

3.4 Information Gain of Diverse Feedback Types

However, when training on a new problem dataset, we lack prior knowledge regarding the type of feedback that will be received for each problem. Additionally, obtaining all feedback types before deciding on the most appropriate one can be costly. Therefore, a preliminary study is conducted to investigate the information gain of each feedback type, which aids in selecting the optimal feedback for each training problem.

For this study, we randomly sample 500 problems that are distinct from the primary problem dataset $\mathcal{D}$, denoted as $\mathcal{D}_{pre}$. We follow the feedback collection procedures outlined in Section 3.2 to gather all three feedback types for these problems. The uncertainty of problem is used as the basis for determining the appropriate feedback type. To assess uncertainty, we employ perplexity $PPL(a, q)$, calculated using the formula:

$$PPL(a, q) = \left(\prod_{i=1}^{l} p(a_i|a_{<i}, q)\right)^{\frac{1}{l}}$$

where l denotes the length of the answer a. For the reward score $R(\cdot|q)$, we use the open-source reward model *OpenAssist-6.9B*, trained on the tasks described earlier. We then plot the relationships between $PPL(a, q)$, $H(a^+, a^-|q)$, $H(a^+, a^-|q, \pi^*)$, and $IG(a^+, a^-|q)$ as scatter plots in three subfigures of Figure 2, respectively. Vertical lines are drawn to evenly divide the problems into three uncertainty groups: low, medium and high.

From the plots, we observe the following: For problems with the lowest uncertainty, critique feedback slightly outperforms the other two feedback types, primarily due to its higher $H(a^+, a^-|q)$. For problems of medium uncertainty, refinement feedback outperforms critique and preference feedback, as evidenced by its lower $H(a^+, a^-|q, \pi^*)$. For the most uncertain problems, preference feedback exhibits the highest information gain, surpassing the other two types, likely due to its ability to sample more informative pairs. These findings provide a solid foundation for developing the feedback combination strategy in our proposed DAIF method.

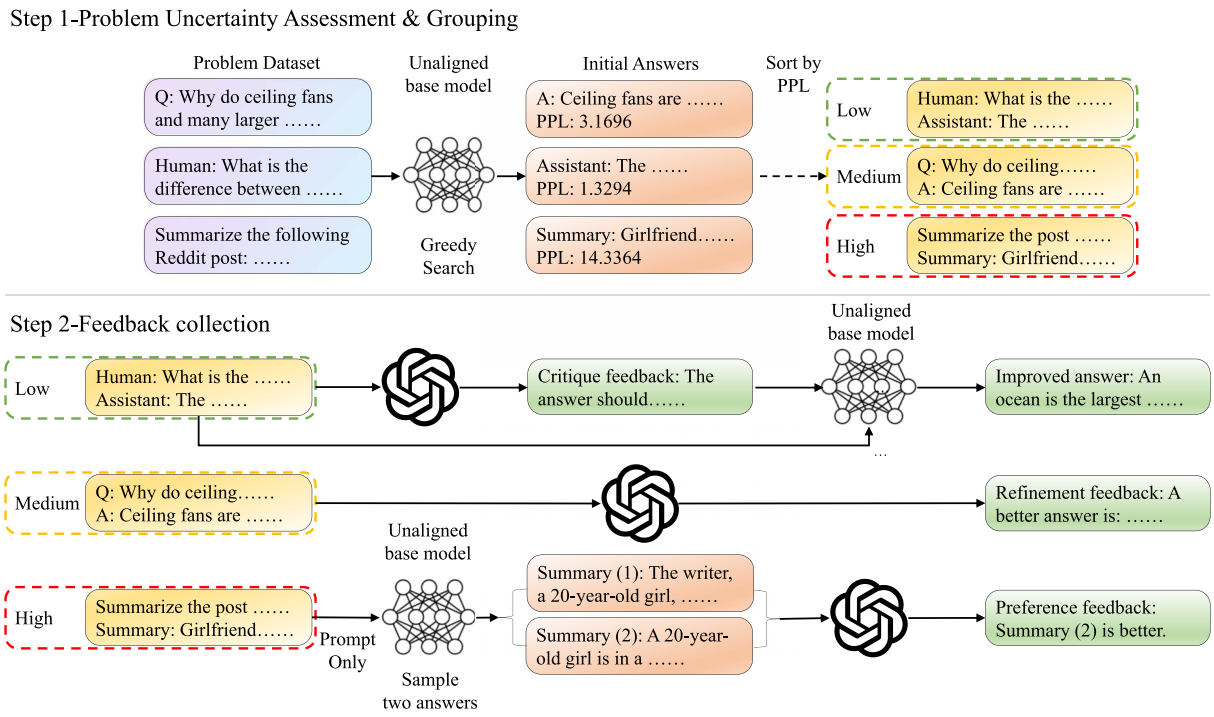
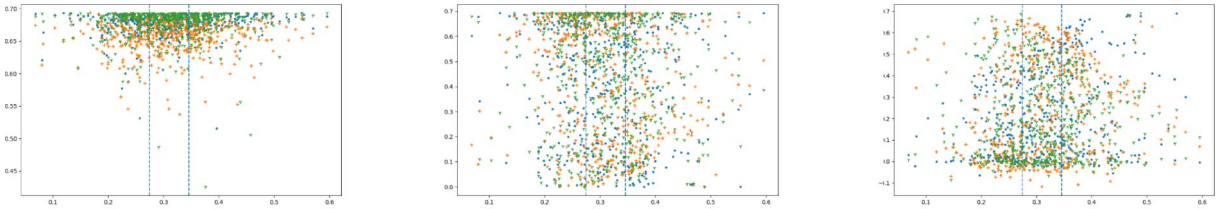


Figure 1: The illustration of our method, DAIF, describing our feedback collection process.



(a) $H(a^+, a^-|q)$ (b) $H(a^+, a^-|q, \pi^*)$ (c) $IG(a^+, a^-|q)$

Figure 2: Scatter plot of $H(a^+, a^-|q)$, $H(a^+, a^-|q, \pi^*)$ and information gain $IG(a^+, a^-|q)$ of every problem in $\mathcal{D}_{pre}$. Blue ''·'' stands for preference feedback, orange ''+'' stands for refinement feedback, and green ''Y'' stands for critique feedback.

4 Method

In this section, we provide a detailed explanation of our proposed DAIF method, which is depicted in Figure 1.

4.1 Uncertainty Assessment and Grouping

To evaluate the uncertainty level of each problem in the dataset $\mathcal{D}$, we first generate an answer a for each problem q in $\mathcal{D}$ using a greedy search algorithm. Then, we calculate the perplexity $PPL(a, q)$ for each answer a. After computing the perplexity score for each problem in $\mathcal{D}$, we first organize the problems in ascending order of perplexity. We then evenly distribute them into three groups: the *low* group $\mathcal{L}$ comprising problems with the lowest perplexity; the *medium* group $\mathcal{M}$ containing problems with moderate perplexity; and the *high* group $\mathcal{H}$ including problems with the highest perplexity scores.

4.2 Feedback Collection

In the next stage of our experiment, we concentrate on collecting diverse feedback for the answers generated by the model. According to the idea of information gain solution and the results discussed in Section 3, we obtain the **Critic** feedback for problems in the *low* group $\mathcal{L}$, the **Refine** feedback for the *medium* group $\mathcal{M}$, and the **Prefer** feedback for the *high* group $\mathcal{H}$. Given the financial and logistical challenges linked to human annotation, we choose to employ the feedback generated by the GPT-3.5-turbo API.

4.3 Training

The final phase of our approach involves model training using the amassed feedback dataset. To demonstrate the effectiveness of our proposed method, we conduct experiments by employing two distinct training methods: Proximal Policy Optimization (PPO) and Direct Preference Optimization (DPO), which are denoted as RM-PPO$_{DAIF}$ and DPO$_{DAIF}$, respectively.

4.3.1 RM-PPO Training

Considering that the supervised fine-tuning process has already been performed on our base model *Vicuna-7B*, we follow the second and third steps of the standard RLHF training procedure. Initially, we train a reward model using the gathered feedback. Subsequently, we fine-tune the policy using the PPO algorithm.

Comparison Dataset Construction. To train a reward model, we initially convert the collected feedback into a comparison format, following Ouyang et al. (2022). This involves the following formats for different types of feedback: (i) for preference feedback, we use the answer pair (a_1, a_2), where the API designates the preferred answer; (ii) for refinement feedback, we adopt the answer pair (a_r, a), where a_r is the improved answer indicated by the API; (iii), for critique feedback, we use the answer pair (a_c, a), where a_c represents the preferred version.

Reward Model Training. Before training the reward model, we partition the feedback dataset into a training set including 90% of the samples and a validation set comprising the remaining 10%. Then, we proceed to train the reward model for five epochs, starting from the base model. The model checkpoint with the highest validation accuracy is chosen for the subsequent PPO training phase.

Policy Model Training. In a manner similar to how we curated the dataset $\mathcal{D}$, we collect an additional 30,000 prompts that are distinct from $\mathcal{D}$ to train the policy model. Due to the limited data available in the *WebGPT-comparison* dataset, we also use alternative sources such as *eli5*, *trivia-qa*, and *ARC* for question-answering prompts. The base model is then trained on these prompts for one epoch using the PPO algorithm. To address concerns related to overfitting, we employ the PPO-ptx strategy Ouyang et al. (2022). The overall training objective can be formally expressed as:

$$\mathcal{J}_{PPO} = \mathbb{E}_{(q,a)\sim\mathcal{D}^{ppo}}\left[r_\theta(q,a) - \beta\log\left(\frac{\pi_{ppo}(a|q)}{\pi_0(a|q)}\right)\right] + \gamma\mathbb{E}_{x\sim\mathcal{D}^{ptx}}\left[\log\pi_{ppo}(x)\right]$$

In this equation, $\mathcal{D}^{ppo}$ represents the dataset of collected prompts, $\mathcal{D}^{ptx}$ signifies the pretraining distribution, π_{ppo} represents the learned PPO policy, and π_0 refers to the base model.

4.3.2 DPO Training

We also train our base model using the DPO algorithm (Rafailov et al., 2023), as DPO offers improved training stability for optimizing the alignment target. DPO is trained on the training split of the comparison dataset, and the overall training objective can be formally expressed as:

$$\mathcal{J}_{DPO} = \mathbb{E}_{(q,a_w,a_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\left(\log\frac{\pi_{dpo}(a_w|q)}{\pi_0(a_w|q)} - \log\frac{\pi_{dpo}(a_l|q)}{\pi_0(a_l|q)}\right)\right)\right] + \gamma\mathbb{E}_{x\sim\mathcal{D}^{ptx}}\left[\log\pi_{dpo}(x)\right]$$

where (a_w, a_l) denotes the answer pair of the problem $q \in \mathcal{D}$ and a_w is the preferred one.

5 Experiments

5.1 Evaluation Settings

We conduct experiments that specifically target three downstream tasks: question answering (QA), dialogue generation (Dial.), and text summarization (Summ.). To assess the effectiveness of our alignment strategy, we construct a separate test set comprising 3,000 prompts. This test set includes 1,000 prompts sampled for each of the three tasks and is entirely distinct from those used in both $\mathcal{D}$ and $\mathcal{D}^{ppo}$. We compare the performance of DAIF against the golden response annotated in the original dataset and our base model (Vicuna-7B). To further demonstrate the benefits of integrating multiple feedback types, we also conduct RM-PPO and DPO training using the datasets restricted to single types of feedback, which are denoted as (RM-PPO$_{Prefer}$, RM-PPO$_{Refine}$, RM-PPO$_{Critic}$) and (DPO$_{Prefer}$, DPO$_{Refine}$, DPO$_{Critic}$), respectively. To accomplish this, we gather critique, refinement, and preference feedback for all problem-answer pairs in $\mathcal{D}$.

Additionally, we carry out comparisons against **Random** settings (RM-PPO$_{Random}$ and

Model	GPT-4 Scoring				RM Scoring			
	Avg.	Summ.	Dial.	QA.	Avg.	Summ.	Dial.	QA.
Golden	2.79	2.98	3.03	2.35	−0.218	0.005	−0.207	−0.452
Vicuna-7B	3.23	3.34	3.08	3.26	0.456	0.526	0.355	0.487
RM-PPO$_{Critic}$	3.58	3.31	3.79	3.64	0.683	0.589	0.741	0.719
RM-PPO$_{Refine}$	3.62	3.27	3.89	3.70	0.708	0.564	0.792	0.767
RM-PPO$_{Prefer}$	3.83	3.74	3.92	3.83	0.731	0.630	0.799	0.765
RM-PPO$_{Random}$	3.69	3.65	3.80	3.63	0.695	0.623	0.754	0.707
RM-PPO$_{Full}$	3.56	3.34	3.78	3.57	0.662	0.592	0.724	0.671
RM-PPO$_{DAIF}$ (Ours)	3.96	4.02	3.94	3.91	0.836	0.891	0.796	0.822
DPO$_{Critic}$	3.68	3.67	3.54	3.83	0.689	0.627	0.633	0.806
DPO$_{Refine}$	3.60	3.44	3.42	3.95	0.694	0.562	0.590	0.930
DPO$_{Prefer}$	3.78	4.11	3.52	3.70	0.798	1.013	0.608	0.772
DPO$_{Random}$	3.69	3.71	3.49	3.88	0.724	0.708	0.613	0.850
DPO$_{Full}$	3.59	3.36	3.51	3.91	0.671	0.582	0.587	0.843
DPO$_{DAIF}$ (Ours)	**4.16**	**4.32**	**3.99**	**4.17**	**1.010**	**1.186**	**0.852**	**0.993**

Table 1: Evaluation results of DAIF and baselines trained with RM-PPO and DPO algorithms in terms of GPT-4 and RM evaluation.

DPO$_{Random}$) and **Full** settings (RM-PPO$_{Full}$ and DPO$_{Full}$), in order to eliminate the influence of different source of chosen answers (Vicuna-7b in **Critic** & **Prefer** feedback and GPT-3.5-Turbo in **Refine** feedback). In **Random** settings, we randomly choose a feedback type from **Critic**, **Refine** and **Prefer** for every prompt without grouping them by uncertainty. In **Full** settings, we adopt all three feedback data of every prompt in $\mathcal{D}$ for training.

5.2 Evaluation Metrics

We present our experimental results using three evaluation metrics: automatic assessment, model-based evaluation, and human-based evaluation. Our primary metric is based on an open-source reward model called OpenAssist,[2] which automatically evaluates the quality of the generated content. We denote this metric as **RM** evaluation. In addition, recent studies have demonstrated the effectiveness of GPT-4 in evaluating chat assistant responses and aligning with human preferences (Zheng et al., 2023; Wang et al., 2023a). Therefore, we incorporate GPT-4 to rate the generated content on a scale from 1 to 5, where higher scores indicate better alignment with human values. We denote this metric as **GPT-4** evaluation. Finally, we acknowledge that human judgment serves as the gold standard for assessing alignment with human values. To address this, we engage 6 NLP researchers to perform pairwise comparisons among the top-performing models identified in automated evaluations. Due to the expensive cost, we only sample 300 test examples (100 for each task) for human evaluation.

[2] https://huggingface.co/OpenAssistant/oasst-rm-2-pythia-6.9b-epoch-1.

5.3 Main Results

Table 1 presents a summary of performance metrics for our DAIF method and the compared benchmarks, evaluated using *OpenAssist* and GPT-4. In both RM-PPO and DPO training scenarios, DAIF consistently outperforms all baseline models in terms of average reward and GPT-4 scores.

As shown in Table 1, DAIF significantly improves alignment performance over the base model *Vicuna-7B* across all tasks. Remarkably, DAIF also exceeds the human-favored answers in the original dataset, demonstrating the effectiveness of using diverse feedback types for training.

When compared to single-feedback approaches, DAIF's combination of multiple feedback types provides a clear advantage, particularly in the **Random** and **Full** settings, where the task-specific feedback tailored to uncertainty levels enhances performance.

Performance analysis within the RM-PPO and DPO paradigms further highlights DAIF's effectiveness, especially in dialog generation under the DPO framework. DPO addresses training instability in PPO, mitigating distribution shifts seen in

		Ours Win (%)	Tie (%)	Ours Lose (%)	Gap (%)
RM-PPO$_{DAIF}$ vs Golden	Dial.	70.0	15.6	14.4	+55.6
	QA.	75.8	13.7	10.5	+65.3
	Summ.	74.4	12.2	13.3	+61.1
	Avg.	**73.5**	13.8	12.7	+**60.8**
RM-PPO$_{DAIF}$ vs RM-PPO$_{Prefer}$	Dial.	22.5	63.0	14.5	+8.0
	QA.	32.3	52.3	15.4	+16.9
	Summ.	46.6	29.6	23.8	+22.8
	Avg.	**33.8**	48.3	17.9	+**15.9**
DPO$_{DAIF}$ vs Golden	Dial.	80.0	11.1	8.9	+71.1
	QA.	76.8	10.5	12.6	+64.2
	Summ.	77.8	8.9	13.3	+64.5
	Avg.	**78.2**	10.2	11.6	+**66.6**
DPO$_{DAIF}$ vs DPO$_{Prefer}$	Dial.	55.8	32.4	11.8	+44.0
	QA.	55.4	23.6	21.0	+34.4
	Summ.	43.1	32.5	25.4	+18.7
	Avg.	**51.4**	29.5	19.1	+**32.3**

Table 2: Human evaluation results. This table shows the performance of RM-PPO$_{DAIF}$ and DPO$_{DAIF}$ against their counterparts across different contexts, showcasing Win, Tie, Lose, and Gap percentages. Here, Gap stands for the difference between Win and Lose percentages.

RM and PPO data, thus further enhancing DAIF's alignment capabilities.

5.4 Human Evaluation

Although automated reward models such as OpenAssist and GPT-4 offer scalability, they possess inherent limitations, including positional and verbosity biases. Consequently, human evaluations play a vital role in accurately gauging alignment with human preferences. To facilitate human annotation processes, our focus shifts to comparing DAIF with key baselines within the RM-PPO and DPO training scenarios:

DAIF vs. Golden: This comparison assesses whether DAIF, benefiting from diverse simulated feedback types, can surpass the human-preferred responses annotated in the original datasets.

DAIF vs. Singular Feedback: This analysis aims to confirm the effectiveness of DAIF's multifaceted feedback approach in comparison to models trained exclusively on a single feedback type. Notably, RM-PPO$_{Prefer}$ and DPO$_{Prefer}$ emerge as the optimal individual feedback types for RM-PPO and DPO training, respectively.

Table 2 presents the human evaluation results, highlighting DAIF's consistent superiority and its distinct advantages. Remarkably, human evaluators rate DAIF's predictions higher than the human-preferred responses from the original datasets in both training scenarios. This underscores DAIF's ability to capture human preferences as reflected in the data effectively. Furthermore, DAIF's dominance over the baseline models trained with a single type of feedback reinforces the hypothesis that leveraging diverse feedback types significantly enhances alignment performance.

In addition, a detailed analysis highlights DAIF's notable performance, especially in text summarization where Vicuna-7B has not received comprehensive pertaining when compared with the models trained on a single feedback type. This underscores DAIF's ability to address tasks that were not extensively covered during the pretraining or supervised fine-tuning phases. However, this trend does not hold when comparing DAIF to human-annotated golden answers, indicating a noticeable difference between human and GPT-4 preferences in the context of text summarization.

6 Analysis

6.1 Over-optimization in RM-PPO Training

Previous studies have acknowledged over-optimization as a common issue in RM-PPO training (Gao et al., 2023; Dubois et al., 2023).

Feedback	Critic	Refine	Prefer	Random	Full	DAIF
Accuracy	86.45	83.22	73.47	80.76	77.95	**88.52**

Table 3: Reward model accuracy of DAIF and baselines in RM-PPO training. The first row stands for the corresponding RM-PPO training setting (e.g., ‘‘Critic’’ represents RM-PPO$_{Critic}$).

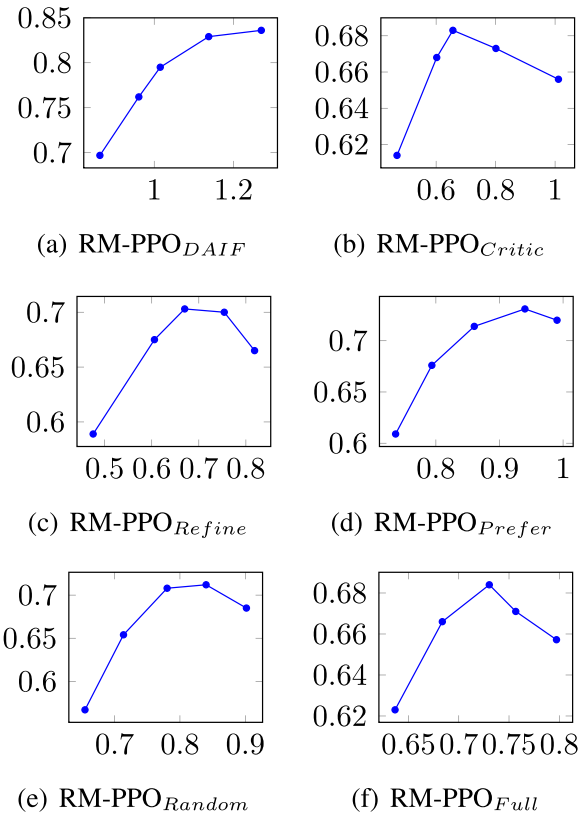
We postulate that DAIF's superior performance, compared to the methods relying on singular feedback types, primarily stems from its capacity to alleviate over-optimization. To evaluate this hypothesis, we conduct additional experiments specifically targeting the aspects of the RM-PPO training process, namely, the reward model and RM-PPO training.

Analysis of the Reward Model. For the reward model, we assess the precision of models trained with DAIF and individual feedback types on a separate test set. This test dataset, comprising 2,700 comparison answer pairs derived from 300 uniquely sampled problems across each task, was structured to assess feedback variations—critique, refinement, preference—on problem answers. As illustrated in Table 3, the reward model of DAIF exhibits superior performance, suggesting its capacity to more comprehensively capture human values in contrast to all other baselines, including models relying on a singular feedback type, randomly selecting feedback types or incorporating all feedback, thus enhancing policy training.

Analysis of RM-PPO Training. In the context of RM-PPO training, we analyze the learning dynamics of various training configurations, illustrating the correlation between proxy reward (x-axis) and OpenAssist evaluation score (y-axis) across PPO iterations in Figure 3. Our findings confirm that our RM-PPO$_{DAIF}$ model effectively optimizes the proxy reward, enhancing the RM score until the onset of over-optimization, after which the RM score changes slightly despite improvements in the proxy reward. In contrast, all other baselines show a notable susceptibility to over-optimization, evident in the occurrence of premature over-optimization points during the PPO training phase.

6.2 Ablation Study

To further investigate the impact of two critical components in our proposed method–grouping strategy and feedback type selection–we conduct



(a) RM-PPO$_{DAIF}$ (b) RM-PPO$_{Critic}$

(c) RM-PPO$_{Refine}$ (d) RM-PPO$_{Prefer}$

(e) RM-PPO$_{Random}$ (f) RM-PPO$_{Full}$

Figure 3: The correlation between RM score (y-axis) and proxy reward (x-axis) across RM-PPO iterations of DAIF and baselines.

ablation studies by modifying the experimental settings for each factor independently.

Grouping Strategy. We explore two variations of the grouping strategy to assess its effectiveness:

1. **Grouping by Task Type:** In the ‘‘By Task’’ condition, we group prompts based on their task type, resulting in three distinct groups: *dialog*, *QA*, and *summarization*. We select **Critique** feedback for the *dialog* tasks with the lowest uncertainty, **Refinement** feedback for *QA* tasks with medium uncertainty, and **Preference** feedback for *summarization* tasks with the highest uncertainty, following the uncertainty distribution presented in Appendix B.
2. **Grouping by Uncertainty within Task Type:** In the ‘‘Within Tasks’’ condition,

Model	RM-PPO results				DPO results			
	Avg.	Summ.	Dial.	QA.	Avg.	Summ.	Dial.	QA.
Ours	**0.836**	**0.891**	**0.796**	**0.822**	**1.010**	**1.186**	**0.852**	**0.993**
By Task	0.775	0.810	0.656	0.805	0.872	0.968	0.785	0.863
Within Tasks	0.754	0.711	0.740	0.809	0.847	0.738	0.824	0.980
By Score	0.769	0.753	0.778	0.775	0.972	1.149	0.825	0.942

Table 4: Evaluation results of DAIF compared to ablation settings on grouping method trained with RM-PPO and DPO algorithms in terms of RM evaluation.

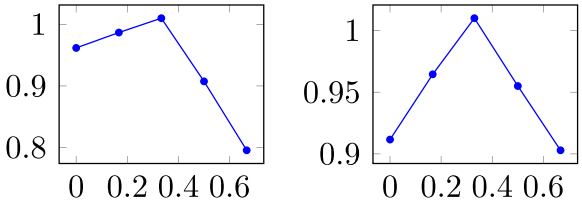
Feedback type			RM-PPO results				DPO results			
$\mathcal{L}$	$\mathcal{M}$	$\mathcal{H}$	Avg.	Summ.	Dial.	QA.	Avg.	Summ.	Dial.	QA.
Critic	Refine	Prefer	**0.836**	**0.891**	**0.796**	0.822	**1.010**	**1.186**	**0.852**	**0.993**
Critic	Prefer	Refine	0.765	0.780	0.767	0.748	0.858	0.898	0.786	0.890
Refine	Critic	Prefer	0.659	0.696	0.482	0.799	0.787	0.941	0.699	0.721
Refine	Prefer	Critic	0.627	0.614	0.511	0.756	0.704	0.611	0.756	0.745
Prefer	Refine	Critic	0.639	0.602	0.482	0.833	0.665	0.591	0.598	0.806
Prefer	Critic	Refine	0.804	0.865	0.707	**0.840**	0.890	0.924	0.809	0.937

Table 5: Evaluation results of DAIF compared to other permutations of feedback types trained with RM-PPO and DPO algorithms in terms of RM evaluation. The first row of the results stands for our proposed method and other rows are ablation settings.

we apply the grouping strategy described in Section 4.1 independently for each specific task. This results in nine distinct groups: $\mathcal{L}_q$, $\mathcal{M}_q$, and $\mathcal{H}_q$ for QA, $\mathcal{L}_d$, $\mathcal{M}_d$, and $\mathcal{H}_d$ for dialog, and $\mathcal{L}_s$, $\mathcal{M}_s$, and $\mathcal{H}_s$ for summarization. We then aggregate the low groups $\mathcal{L}_q$, $\mathcal{L}_d$, and $\mathcal{L}_s$ into an overall *low* group $\mathcal{L}$, and similarly aggregate the medium and high groups into *medium* ($\mathcal{M}$) and *high* ($\mathcal{H}$) groups, respectively.

We present the results of our reward model scoring for both the proposed method and the ablation variants in Table 4. The findings indicate that our method outperforms both ablation settings, supporting the following conclusions: (1) Uncertainty, rather than task type, serves as a more reliable criterion for grouping, aligning with human learning paradigms. (2) It is unnecessary to consider the specific task type when determining the uncertainty group to which a problem should be assigned.

Feedback Type Selection Method. To validate the efficacy of our feedback type selection strategy based on information gain, we compare all possible permutations of the three feedback types



(a) Fix Critique (b) Fix Preference

Figure 4: Experimental results on varying the proportion of different feedback types. The horizontal axes in both subfigures denote the proportion of refinement feedback.

(Critique, Refinement, and Preference) within different uncertainty groups. The results, shown in Table 5, highlight the superiority of our feedback combination strategy. By maximizing information gain through the optimal combination of diverse feedback types, our method outperforms the other strategies, achieving better preference scores according to the reward model.

6.3 The Proportion of Different Feedback Types

To further evaluate the effectiveness of different feedback types, we conducted experiments by varying the proportions of each type of feedback.

Case #1: Low	
Input Prompt	*Human*: Can you describe what happened in the war of 1812 *Assistant*: Sure, that's a pretty good story. The war was very unpopular in the United States. *Human*: tell me about it Assistant: .....
Initial Answer a	The war ended in 1815 with the signing of the Treaty of Ghent, which was a peace treaty that ended the war. The treaty did not address the issues that had caused the war, but it did establish peace between the two countries. (RM score: -0.2739)
Critique Feedback	To improve the response, the assistant could provide more details about the war, such as the major battles, the reasons for the conflict, and the consequences .....
Improved Answer a_c	The War of 1812 was a conflict between the United States and Great Britain that lasted from 1812 to 1815. The war was largely fought in the United States and ended with the signing of the Treaty of Ghent ..... (RM score: 0.6846)
Refinement Feedback (a_r)	The War of 1812 ended in 1815 with the signing of the Treaty of Ghent, which was a peace treaty between the United States and Britain. The treaty .....(RM score: 0.5166)
Answer a'	The war was very unpopular in the United States, and many people wanted to end it. The Treaty of Ghent was ..... (RM score: -0.8306)
Preference Feedback	The initial answer a is better.
Case #2: Medium	
Input Prompt	*Human*: What happens in our brains when we feel inferior, even when we consciously know we shouldn't? For example, ...
Initial Answer a	It's possible that when we feel inferior, our brains release certain chemicals that trigger certain responses, such as .....(RM score: 0.5674)
Critique Feedback	There are a few areas where improvements can be made: 1. Provide more specific information: .....
Improved Answer a_c	It's possible that when we feel inferior, our brains release certain chemicals that trigger certain responses, such as avoidance behaviors. This could be ..... (RM score: 1.7549)
Refinement Feedback (a_r)	When we feel inferior, our brains may release certain chemicals that trigger specific responses, such as avoidance behaviors. This can be attributed to .....(RM score: 2.7188)
Answer a'	When we feel inferior, our brains can trigger a response known as the ''fight or flight'' response. This response is ..... (RM score: 1.6504)
Preference Feedback	Answer a' is better.
Case #3: High	
Input Prompt	*Human*: Please summarize the following Reddit post in no more than 100 words. SUBREDDIT: r/relationships TITLE: Me with my girlfriend 4years, she doesn't want relationship because she is confused POST: hi everyone I need help from everyone who is reading this..... *Assistant*: *TL;DR*:
Initial Answer a	Boyfriend's girlfriend is confused and doesn't want a relationship, causing him to be depressed and unsure of what to do. (RM score: -1.3057)
Critique Feedback	The intended summary effectively condenses the post, but it could be improved by removing redundant information and clarifying the main conflict.
Improved Answer a_c	A boyfriend is depressed due to his girlfriend's confusion about their relationship and her feelings for a close friend. She wants to be just friends, and he is unsure of how to handle the situation. (RM score: -1.1826)
Refinement Feedback (a_r)	The boyfriend's girlfriend is unsure about their relationship and wants to be just friends due to feelings for a close friend. The boyfriend is depressed and seeking advice on how to handle the situation and not lose her.(RM score: -0.7271)
Answer a'	The author is in a 4-year relationship with their girlfriend, but she is unsure about the future of their relationship due to her own confusion and .....(RM score: 0.3186)
Preference Feedback	Answer a' is better.

Table 6: Three cases chosen from the training set, with one from each uncertainty group.

Specifically, we implemented two experimental settings: (1) maintaining a constant proportion of critique feedback while gradually varying the proportions of refinement and preference feedback, and (2) maintaining a constant proportion of preference feedback while gradually varying the proportions of critique and refinement feedback. Due to space limitations, we report only the average RM score from the DPO training algorithm and present the results in Figure 4. The results indicate that our strategy, with a refinement feedback proportion of approximately 0.333, is closest to

the optimal compared to other proportions tested. This demonstrates that an optimal balance of each feedback type enhances overall performance in our experimental settings.

6.4 Case Analysis

To verify our motivation about the performances of different feedback types, we conduct nuanced case analyses, focusing on different feedback types corresponding to problems of varying uncertainty levels. We have chosen one example from each of the *low*, *medium*, and *high* groups. The specifics of each example include the original problem q, its original answer a, along with the critique, refinement, and preference feedback, which are provided in Table 6.

Generally, we have the following observations. (1) A detailed analysis of the least challenging problems under critique feedback indicates that our model provides a more effective generation of improved answers a_c compared to refined feedback a_s. However, the cases also show that as the uncertainty increases, it becomes apparent that the refined answers a_s progressively outshine the improved answers a_c. This suggests the crossing of a threshold in self-improvement capability as the uncertainty level escalates. (2) On the contrary, the comparative analysis of the answer a and the answer a' from preference feedback accentuates its increasing relevance with the escalation of problem uncertainty. As shown by the cases in Table 6, while the solutions for easier problems tend to be largely similar, the disparity in quality becomes evident as the complexity of problems increases. This highlights the distinct advantage of incorporating annotated preferences for model learning, particularly in more challenging scenarios. (3) These instances validate our initial analyses and underscore the rationale behind our strategic feedback approach. They emphasize the adaptability and effectiveness of our method in aligning LLMs with varying levels of problem uncertainty.

7 Conclusion

In this paper, we introduced DAIF, an innovative data collection methodology designed to improve the alignment of LLMs with human values. We proposed a differentiated feedback approach, where various types of feedback are utilized based on the varying uncertainty levels of the problems. We presented experimental results across three downstream tasks, demonstrating that DAIF achieved superior performance even with a smaller dataset. Further experiments and analyses provided additional evidence of the effectiveness of DAIF. We believe that our method opens up new possibilities for the effective utilization of feedback in aligning LLMs.

Acknowledgments

Min Yang was supported by the National Key Research and Development Program of China (2022YFF0902100), the National Natural Science Foundation of China (grant no. 62376262), the Natural Science Foundation of Guangdong Province of China (2024A1515030166), Shenzhen Science and Technology Innovation Program (KQTD20190929172835662), and the Alibaba Innovative Research Program.

References

Afra Feyza Akyürek, Ekin Akyürek, Aman Madaan, Ashwin Kalyan, Peter Clark, Derry Wijaya, and Niket Tandon. 2023. Rl4f: Generating natural language feedback with reinforcement learning for repairing model outputs. *arXiv preprint arXiv:2305.08844*. https://doi.org/10.18653/v1/2023.acl-long.427

Joris Baan, Nico Daheim, Evgenia Ilia, Dennis Ulmer, Haau-Sing Li, Raquel Fernández, Barbara Plank, Rico Sennrich, Chrysoula Zerva, and Wilker Aziz. 2023. Uncertainty in natural language generation: From theory to applications. *arXiv preprint arXiv:2307.15703*.

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. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

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. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Erdem Bıyık, Dylan P. Losey, Malayandi Palan, Nicholas C. Landolfi, Gleb Shevchuk, and Dorsa Sadigh. 2022. Learning reward functions from diverse sources of human feedback: Optimally integrating demonstrations and preferences. *The International Journal of Robotics Research*, 41(1):45–67. https://doi.org/10.1177/02783649211041652

Erdem Bıyık, Malayandi Palan, Nicholas C. Landolfi, Dylan P. Losey, and Dorsa Sadigh. 2019. Asking easy questions: A user-friendly approach to active reward learning. *arXiv preprint arXiv:1910.04365*.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4):324–345. https://doi.org/10.1093/biomet/39.3-4.324

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*.

Nirjhar Das, Souradip Chakraborty, Aldo Pacchiano, and Sayak Ray Chowdhury. 2024. Active preference optimization for sample efficient rlhf. In *ICML 2024 Workshop on Theoretical Foundations of Foundation Models*.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback. *arXiv preprint arXiv:2305.14387*.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

Huawen Feng, Yan Fan, Xiong Liu, Ting-En Lin, Zekun Yao, Yuchuan Wu, Fei Huang, Yongbin Li, and Qianli Ma. 2024. Improving factual consistency of news summarization by contrastive preference optimization. In *Findings of EMNLP 2024*, pages 11084–11100. https://doi.org/10.18653/v1/2024.findings-emnlp.648

Ge Gao, Eunsol Choi, and Yoav Artzi. 2022. Simulating bandit learning from user feedback for extractive question answering. *arXiv preprint arXiv:2203.10079*. https://doi.org/10.18653/v1/2022.acl-long.355

Haoyu Gao, Ting-En Lin, Hangyu Li, Min Yang, Yuchuan Wu, Wentao Ma, Fei Huang, and Yongbin Li. 2024. Self-explanation prompting improves dialogue understanding in large language models. In *LREC-COLING 2024*, pages 14567–14578.

Leo Gao, John Schulman, and Jacob Hilton. 2023. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR.

Gaurav R. Ghosal, Matthew Zurek, Daniel S. Brown, and Anca D. Dragan. 2023. The effect of modeling human rationality level on learning rewards from multiple feedback types. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5983–5992. https://doi.org/10.1609/aaai.v37i5.25740

Adam Gleave and Geoffrey Irving. 2022. Uncertainty estimation for language reward models. *arXiv preprint arXiv:2203.07472*.

Mengting Hu, Zhen Zhang, Shiwan Zhao, Minlie Huang, and Bingzhe Wu. 2023. Uncertainty in natural language processing: Sources, quantification, and applications. *arXiv preprint arXiv:2306.04459*.

Hong Jun Jeon, Smitha Milli, and Anca Dragan. 2020. Reward-rational (implicit) choice: A unifying formalism for reward learning. *Advances in Neural Information Processing Systems*, 33:4415–4426.

Seungjae Jung, Gunsoo Han, Daniel Wontae Nam, and Kyoung-Woon On. 2024. Binary classifier optimization for large language model alignment. *arXiv preprint arXiv:2404.04656*.

Yann LeCun, Sumit Chopra, Raia Hadsell, M. Ranzato, Fujie Huang. 2006. A tutorial on energy-based learning. In *Predicting Structured Data*. https://doi.org/10.7551/mitpress/7443.003.0014

Minghao Li, Feifan Song, Bowen Yu, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. 2023. Api-bank: A benchmark for tool-augmented llms. *arXiv preprint arXiv:2304.08244*.

Hao Liu, Carmelo Sferrazza, and Pieter Abbeel. 2023a. Chain of hindsight aligns language models with feedback. *arXiv preprint arXiv:2302.02676v6*, 3.

Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Denny Zhou, Andrew M. Dai, Diyi Yang, and Soroush Vosoughi. 2023b. Training socially aligned language models in simulated human society. *arXiv preprint arXiv:2305.16960*.

Run Luo, Haonan Zhang, Longze Chen, Ting-En Lin, Xiong Liu, Yuchuan Wu, Min Yang, Minzheng Wang, Pengpeng Zeng, Lianli Gao, Heng Tao Shen, Yunshui Li, Xiaobo Xia, Fei Huang, Jingkuan Song, and Yongbin Li. 2024. Mmevol: Empowering multimodal large language models with evol-instruct. *arXiv preprint arXiv:2409.05840*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.

Viraj Mehta, Vikramjeet Das, Ojash Neopane, Yijia Dai, Ilija Bogunovic, Jeff Schneider, and Willie Neiswanger. 2023. Sample efficient reinforcement learning from human feedback via active exploration. *arXiv preprint arXiv:2312.00267*.

Luckeciano C. Melo, Panagiotis Tigas, Alessandro Abate, and Yarin Gal. 2024. Deep bayesian active learning for preference modeling in large language models. *arXiv preprint arXiv:2406.10023*.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. 2023. Scaling data-constrained language models. *arXiv preprint arXiv:2305.16264*.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll 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. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Malayandi Palan, Nicholas C. Landolfi, Gleb Shevchuk, and Dorsa Sadigh. 2019. Learning reward functions by integrating human demonstrations and preferences. *arXiv preprint arXiv:1906.08928*. https://doi.org/10.15607/RSS.2019.XV.023

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*.

Jérémy Scheurer, Jon Ander Campos, Jun Shern Chan, Angelica Chen, Kyunghyun Cho, and Ethan Perez. 2022. Training language models with natural language feedback. *arXiv preprint arXiv:2204.14146*, 8.

Burr Settles. 2009. Active learning literature survey. https://burrsettles.com/pub/settles.activelearning.pdf

Weiyan Shi, Emily Dinan, Kurt Shuster, Jason Weston, and Jing Xu. 2022. When life gives you lemons, make cherryade: Converting feedback from bad responses into good labels. *arXiv preprint arXiv:2210.15893*.

Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. 2023. Preference ranking optimization for human alignment. *arXiv preprint arXiv:2306.17492*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Niket Tandon, Aman Madaan, Peter Clark, and Yiming Yang. 2021. Learning to repair: Repairing model output errors after deployment using a dynamic memory of feedback. *arXiv preprint arXiv:2112.09737*. https://doi.org/10.18653/v1/2022.findings-naacl.26

Zhengwei Tao, Ting-En Lin, Xiancai Chen, Hangyu Li, Yuchuan Wu, Yongbin Li, Zhi Jin, Fei Huang, Dacheng Tao, and Jingren Zhou. 2024. A survey on self-evolution of large language models. *arXiv preprint arXiv:2404.14387*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. 2022. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*.

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023a. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

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

Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi. 2022. Generating sequences by learning to self-correct. *arXiv preprint arXiv:2211.00053*.

Yue Wu, Zhiqing Sun, Huizhuo Yuan, Kaixuan Ji, Yiming Yang, and Quanquan Gu. 2024. Self-play preference optimization for language model alignment. *arXiv preprint arXiv:2405.00675*.

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Banghua Zhu, Jiantao Jiao, and Michael I. Jordan. 2023. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. *arXiv preprint arXiv:2301.11270*.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Analysis of Experimental Settings

In this experiment, we use uncertainty as the basis for grouping and feedback type selection. Alternative metrics, such as the base model's task performance, could also be applied, introducing potential variability. To assess the impact of the chosen uncertainty metric, we conduct a comparative analysis of DAIF's performance using the reward score of an open-source model as an alternative uncertainty indicator.

In the ''By Score'' condition, we deviate from the conventional approach of evaluating problem uncertainty, $PPL(a, q)$, and instead employ the *OpenAssist-6.9B* reward model to assign a reward score to the answer a in response to the question q. We interpret the negative value of this reward score as the indicator for grouping. The results of the comparison between our method and the ''By Score'' settings, as presented in Table 4, demonstrate that perplexity outperforms the open-source reward model as the basis for grouping and feedback type selection.

Task	Summ.	Dial.	QA.
$\mathcal{L}$	12	6,225	3,763
$\mathcal{M}$	790	3,076	6,128
$\mathcal{H}$	9,198	693	109

Table 7: The number of problems from different tasks in every uncertainty group.

B Distribution of Uncertainty Across Different Tasks

Table 7 presents the distribution of examples across the different uncertainty groups for each task. The data clearly indicates that, for our base model *Vicuna-7B*, the summarization task exhibits the highest average uncertainty, while dialog generation is identified as the easiest task.