

# Aligning Large Language Models with Human Preferences through Representation Engineering

Wenhao Liu<sup>†</sup>, Xiaohua Wang<sup>†</sup>, Muling Wu, Tianlong Li, Changze Lv  
Zixuan Ling, Jianhao Zhu, Cenyuan Zhang, Xiaoqing Zheng\*, Xuanjing Huang

School of Computer Science, Fudan University, Shanghai, China

{whliu22, xiaohuawang22}@m.fudan.edu.cn

{zhengxq, xjhuang}@fudan.edu.cn

## Abstract

Aligning large language models (LLMs) with human preferences is crucial for enhancing their utility in terms of helpfulness, truthfulness, safety, harmlessness, and interestingness. Existing methods for achieving this alignment often involve employing reinforcement learning from human feedback (RLHF) to fine-tune LLMs based on human labels assessing the relative quality of model responses. Nevertheless, RLHF is susceptible to instability during fine-tuning and presents challenges in implementation. Drawing inspiration from the emerging field of representation engineering (RepE), this study aims to identify relevant representations for high-level human preferences embedded in patterns of activity within an LLM and achieve precise control of model behavior by transforming its representations. This novel approach, denoted as Representation Alignment from Human Feedback (RAHF), proves to be effective, computationally efficient, and easy to implement. Extensive experiments demonstrate the efficacy of RAHF in not only capturing but also manipulating representations to align with a broad spectrum of human preferences or values, rather than being confined to a singular concept or function (e.g. honesty or bias). RAHF's versatility in accommodating diverse human preferences shows its potential for advancing LLM performance.

## 1 Introduction

While large language models (LLMs) learn broad-ranging world knowledge and a degree of reasoning proficiency, precise control over their behavior proves challenging due to the unsupervised nature of their pre-training (Radford et al., 2018, 2019; Brown et al., 2020; Bubeck et al., 2023; Touvron et al., 2023). For each query, instruction-tuned LLMs (Wei et al., 2021; Chung et al., 2022; Tou-

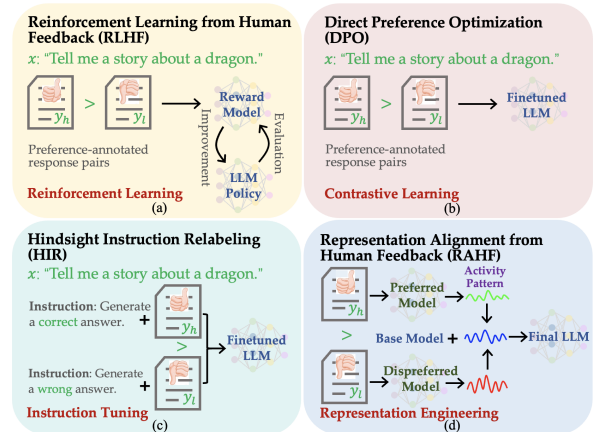


Figure 1: Illustration of different approaches. (a) Reinforcement learning from human feedback (RLHF); (b) Direct preference optimization (DPO); (c) Hindsight instruction relabeling (HIR); (d) Representation alignment from human feedback (RAHF).

vron et al., 2023) exhibit the capacity to generate multiple responses that are both semantically and syntactically coherent by some sampling techniques. While such ability enables the models to provide diversity that is essential for chat agents, some responses may contain harmful, unethical, socially biased, and negative, even illegal content (Srivastava et al., 2022; Thoppilan et al., 2022; Bubeck et al., 2023; Wang et al., 2023).

Existing methods steer LLMs to align with human preferences often using reinforcement learning (RL), with reinforcement learning from human feedback (RLHF) emerging as the most successful one (Ouyang et al., 2022). However, the underlying learning algorithms exhibit a considerable degree of complexity, sensitivity to hyperparameters, instability during training, and necessitate additional training in the reward model and value network, leading to substantial computational costs (Yuan et al., 2023; Rafailov et al., 2023).

In addressing the aforementioned challenges posed by RL-based methods, several computationally lightweight alternatives have been proposed to simplify the human preference-matching process.

<sup>†</sup> These authors contributed equally.

\* Corresponding author.

Two prominent paradigms among these alternatives include contrastive learning (Rafailov et al., 2023; Zhao et al., 2023; Yuan et al., 2023) and Hindsight instruction relabeling (HIR) (Zhang et al., 2023; Liu et al., 2023). Contrastive learning-based methods optimize a language model policy by increasing the relative probability of preferred responses over dispreferred ones, while HIR methods transform human feedback into instructions by relabeling the original ones, indicating the relative quality of provided responses. A common characteristic shared by these two paradigms is their capability to align language models with human preferences through reward-free fine-tuning.

However, the reward-free fine-tuning is vulnerable to the presence of noisy data or incorrect labels in a training set comprising a collection of preference-annotated response pairs (Li et al., 2023b; Dumoulin et al., 2023). Instances of dull sentences or very brief responses may appear repeatedly in such a training set, potentially introducing bias into the models. The exclusion of such instances from the training set renders it impossible for LLMs to glean insights into human preferences expressed in these instances. In contrast, RL-based methods adopt a different strategy, wherein a reward function is first extracted from a dataset of response rankings, and then this reward function can be applied to train an LLM, effectively mitigating the model’s direct exposure to noisy data or incorrect labels within the dataset.

In this study, we aim to seek for a computationally lighter and reward-free algorithm that can effectively harness human preference expressed in datasets meanwhile safeguarding LLMs from the influence of noisy data. Inspired by the recent advance in representation engineering (Zou et al., 2023), we initially locate relevant representations and activity patterns associated with high-level human preferences within an LLM, and subsequently gain precise control over its behavior by manipulating its internal representations. In the neural architecture, network weights determine neural activity, neural activity determines the networks’ output, and the networks’ output determines the networks’ behavior. Instead of focusing on neurons and their connections, we see aligning LLMs with human feedback as an outcome of representational spaces, implemented by patterns of activity across populations of neurons. We first identify the differences in model activities between preferred and dispreferred stimuli, and then control its behavior

by leveraging the identified differences in representations (see Figure 1). We introduce two methods for controlling representations and demonstrate the efficacy of these representation engineering (RepE) approaches in aligning LLMs with a broad spectrum of human preferences through a collection of response pairs.

To validate the effectiveness of our approach in aligning with human preferences, we conducted extensive comparative experiments on the generated results. Our method outperformed RLHF and other RL-free approaches in human evaluations and automated metrics such as general abilities and GPT-4 evaluations. Notably, the underlying algorithms exhibit simplicity in implementation and straightforwardness in training.

## 2 Related Work

Tuning large language models to elicit desired responses and behavior from their extensive knowledge and capabilities is essential in the development of chat agents, such as ChatGPT (Brown et al., 2020), LLaMA (Touvron et al., 2023) and GPT-4 (Bubeck et al., 2023), characterized by safety, performance, and controllability. The enlargement of the size of language models only does not inherently enhance their ability to follow a user’s intent. For example, LLMs may still generate outputs that are untruthful, toxic, or simply not helpful to the user. Existing human preference alignment methods can be broadly classified into three major categories: reinforcement learning (Ouyang et al., 2022; Ramamurthy et al., 2023), contrastive learning (Rafailov et al., 2023; Zhao et al., 2023; Yuan et al., 2023), and Hindsight instruction relabeling (Zhang et al., 2023; Liu et al., 2023).

Extensive research has been devoted to the exploration of RL from human feedback through ratings or rankings, spanning tasks from NL-to-SQL conversion (Zhong et al., 2017), machine translation (Kreutzer et al., 2018), task-oriented dialogue systems (Su et al., 2019; Zhang et al., 2019; Takanobu et al., 2019), summarization (Stiennon et al., 2020), story-telling (Ziegler et al., 2019) to instruction-following (Ouyang et al., 2022; Ramamurthy et al., 2023). Typically, these methods involve the fitting of a reward model to a dataset of human preferences, followed by the optimization of a LLM policy to generate responses with high reward, using RL algorithms such as REINFORCE (Williams, 1992) or proximal policy optimization (Schulman

et al., 2017). Despite the attractiveness of leveraging human preferences that are easier to collect than expert demonstrations, training LLMs with RL poses significant practical challenges, which is attributed to the sensitivity of RL to hyperparameters and the inherent instability during training.

The solutions based on Hindsight instruction re-labeling (Zhang et al., 2023; Liu et al., 2023) and contrastive learning (Rafailov et al., 2023; Zhao et al., 2023; Yuan et al., 2023) have emerged as computationally efficient alternatives to RL-based methods without explicit reward modeling. However, these reward-free fine-tuning solutions are susceptible to noisy data or incorrect labels within a training set. They exhibit performance lags compared to models tuned with RL counterparts (see Section 4). Furthermore, the question of whether LLMs trained with such fine-tuning methods can generalize well to out-of-distribution queries remains unresolved when contrasted with models incorporating an explicit reward model. RLHF method (Ouyang et al., 2022) offers a potential avenue for improvement by leveraging additional unlabeled examples through labeling LLM generations with the learned reward model.

To enhance transparency and controllability of neural networks, Zou et al. (2023) introduced representation engineering (RepE) as a methodology, drawing an analogy between understanding deep neural networks through representation tomography and studying brains via neuroimaging techniques. Their work demonstrated the efficacy of RepE in addressing diverse safety-related challenges such as truthfulness, honesty, and hallucination. This study falls in line with recent research findings and extends its application to aligning LLMs with a wide spectrum of human preferences. Our study introduces two novel methods to instruct LLMs on human preferences first, and then extract differences in model activities between preferred and dispreferred stimuli. These differences in activity patterns serve as a foundation for manipulating the model’s behavior, leading to the generation of responses that better align with human preferences. Due to the lightweight computational advantages of parameter-efficient fine-tuning techniques (Houlsby et al., 2019; Lester et al., 2021; Hu et al., 2021; Wu et al., 2023, 2024), these techniques are utilized to fit the disparity in activity patterns. In contrast to the approach adopted by Zou et al. (2023), which relies on unlabeled or self-generated stimuli limited to singular concepts or functions the meaning

of which the models have already “known”, our methods provide a more comprehensive alignment with diverse human preferences.

### 3 Method

We begin by instructing LLMs on human preferences with a set of preference-annotated response pairs. We introduce two novel methods for instructing LLMs on human preferences and extracting their activity patterns: one involving a single LLM (trained to discern the relative quality of responses) and the other employing dual LLMs (“a good guy and a bad guy”). Secondly, we collect the activity patterns of LLMs when exposed to stimuli that are preferred or dispreferred. The differences in these patterns serve as the foundation for manipulating LLMs, enabling them to generate responses more closely aligned with human values. Finally, we construct the final model by training a low-rank adapter (Hu et al., 2021) to fit the disparity in activity patterns.

#### 3.1 Instructing LLMs on Human Preferences

To extract activity patterns from the model that align with human preferences, it is crucial for the model to possess a correct understanding and awareness of these preferences. The effectiveness of extracting activity patterns from alignment fine-tuned models, such as LLaMA-2-chat, in capturing concepts like truthfulness and honesty has been validated by Zou et al. (2023). However, for non-aligned models, such as pre-trained large language models or LLMs subjected to simple fine-tuning, explicit indications of human preferences should be provided to elicit and capture activity patterns induced by stimulus preferences. This capability enables the accumulation of diverse activities, subsequently utilized to calibrate LLMs based on human preferences.

For instructing LLMs on human preferences, we rely on a dataset annotated with human preferences. As mentioned earlier, we employ two methods to achieve this goal. The first method utilizes Hindsight (Zhang et al., 2023), using contrastive instructions to instruct a single LLM. The second method involves fine-tuning two LLMs separately: one (referred to as the preferred model) is fine-tuned based on preferred responses, while the other (referred to as the dispreferred model) is fine-tuned on dispreferred responses.

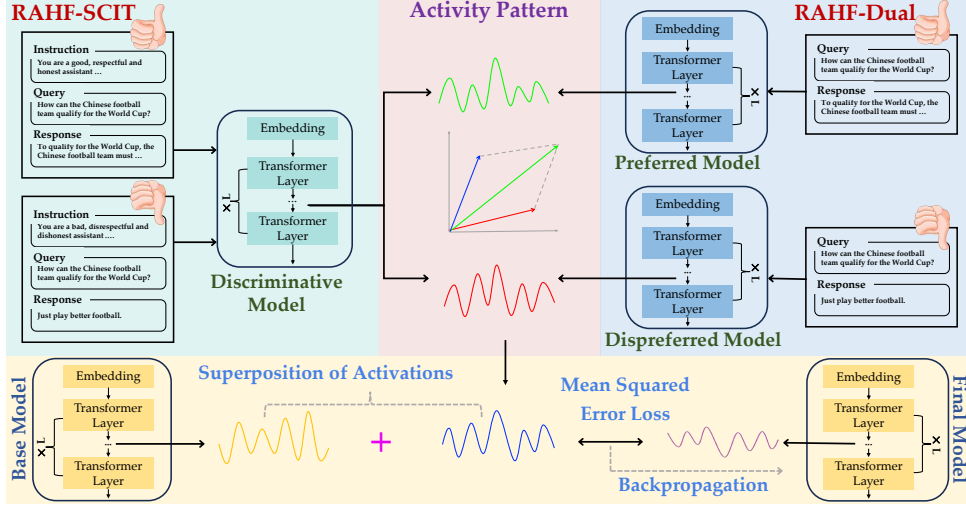


Figure 2: The procedure of RAHF. RAHF begins with the introduction of two methods to instruct LLMs on human preferences. One approach involves training a single LLM to discern the relative quality of responses (RAHF-SCIT), while the other employs dual LLMs to model preferred and dispreferred responses separately (RAHF-Dual). Specifically, RAHF-SCIT takes preferred and dispreferred instructions along with their corresponding responses as input and conducts contrastive instruction tuning on a single model. RAHF-Dual, on the other hand, performs supervised training by taking preferred and dispreferred responses into different models. Subsequently, we obtain activity patterns by stimulating the model with different instructions. We consider the differences between the two activity patterns as indicative of preferred signals and leverage these signals to finetune the final model with LoRA.

### 3.1.1 Preference Instruction with a Single Model

Within the proposed framework, the Single LLM Method focuses on fine-tuning a Single Large Language Model through Contrastive Instruction Tuning (SCIT). This process involves two instructions: one instructs the model to generate responses preferred by humans, while the other guides the model to generate responses dispreferred by humans. Following such fine-tuning, we can optimize the model for consistency with human preferences. We can also stimulate the model to elicit distinct activity patterns by employing different instructions subsequently.

Specifically, the training dataset is curated to include pairs of both preferred and dispreferred instructions, alongside associated queries and their corresponding responses (details on preferred instructions can be found in Appendix A.1). Following HIR(Zhang et al., 2023), for instructions linked to positive preferences, the fine-tuning objective aims to increase the probability of generating preferred responses while concurrently decreasing the probability of generating dispreferred responses. Conversely, for instructions associated with negative preferences, the objective is to elevate the probability of generating dispreferred responses and reduce the probability of generating preferred responses.

Formally, let  $D$  represent the training dataset, with  $q_i$  denoting the query,  $r_i$  representing the response, and  $p_i$  indicating the instruction (positive or negative). The fine-tuning of the LLM involves minimizing the following loss:

$$\mathcal{L} = - \sum_{(p_i, q_i, r_i) \in D} (P^+ + \log \frac{\exp(P^+)}{\exp(P^+) + \exp(P^-)}) \quad (1)$$

where  $P^+ = \log \pi(r_i | p_i, q_i; \theta)$ ,  $P^- = \log \pi(r_i | p_i^*, q_i; \theta)$  and  $p_i^*$  denotes the opposite instruction, ensuring a contrast between preferred and dispreferred cases.

Throughout the entire fine-tuning process, the LLM undergoes a learning phase to distinguish between preferred and non-preferred responses, revealing distinct activity patterns associated with human preferences. Subsequently, these two instructions will serve as stimuli to acquire the model’s internal representations, which will be used for further alignment. This contrastive training relying on preference data enables the achievement of the overarching goal of consistency with a broad spectrum of human preferences, rather than a singular concept.

### 3.1.2 Preference Instruction with Dual Models

In the Dual LLMs method, we aim to train two LLMs with distinct tendencies: one model is inclined to generate preferred responses, while the



other tends to produce dispreferred responses. To achieve this objective, we employ paired preference data to conduct supervised fine-tuning of the LLMs. Specifically, we use the preferred data from the preference pairs to train the preferred model and the dispreferred data from the preference pairs to train the dispreferred model.

Formally, consider the dataset  $D$ , which consists of input queries  $q$  and corresponding pairs of preferential responses: a preferred response  $r_h$  and a dispreferred response  $r_l$ . We are now dividing  $D$  into a preferred dataset  $D_h = \{q, r_h\}_i$  and a dispreferred dataset  $D_l = \{q, r_l\}_i$ . Utilizing this data, we employ a supervised learning approach (maximum likelihood) to fine-tune the LLMs, thereby obtaining two models expressing preferences, denoted as  $\pi_h$  and  $\pi_l$  respectively. The fine-tuning of these two LLMs is aimed at maximizing the achievement of the following objectives:

$$\pi_h(\theta^*) = \arg \max_{\theta} \sum_{(q_i, r_i) \in D_h} \log \pi(r_i | q_i; \theta) \quad (2)$$

$$\pi_l(\theta^*) = \arg \max_{\theta} \sum_{(q_i, r_i) \in D_l} \log \pi(r_i | q_i; \theta) \quad (3)$$

Through this training process, the preferred model and dispreferred model have respectively learned the activity patterns associated with human-preferred and dispreferred responses. Due to the human preference learning conducted in two distinct models, in contrast to SCIT, the Dual LLMs method does not require additional distinct instructions during fine-tuning. Instead, guidance for the model is provided solely through different responses.

### 3.2 Collecting Activity Patterns

Following the establishment of comprehension of human preferences by LLMs, we are able to extract representations of what humans prefer and disprefer. Due to the characteristics of autoregressive Transformer language models, the attention mechanism results in tokens at different positions exhibiting distinct representations. The activation representation of a token at the current position is influenced by preceding tokens. Therefore, for a specific pair of query  $q$  and response  $r$ , this pair is concatenated with two instructions from Section 3.1, which guide the model in forming the concept of human preferences and inputted into the model to obtain the intermediate layer hidden states at each position as internal representations.

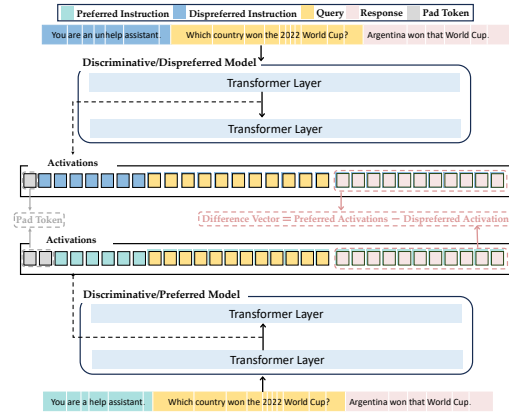


Figure 3: Examples of Collecting Activity Patterns. To ensure the correspondence between the positions of preferred and dispreferred instructions during the extraction of difference vectors, instruction  $p$  and query  $q$  are left-padded to the maximum prompt length, while the response  $r$  is right-padded to the maximum response length.

Formally, for a given instruction  $p$ , a decoder model  $\pi$ , we collect the  $l$ -th layer’s hidden states of each token for the query-response pair  $(q_i, r_i)$  within dataset  $D$ . This can be formalized as follows:

$$A_{p, \pi, l} = \pi_l(p, q_i, r_i) \mid (q_i, r_i) \in D \quad (4)$$

Here,  $\pi_l$  represents the hidden states output by the neural network’s  $l^{th}$  layer. We directly extract the hidden layer states from the neural network as representations. To address the issue of varying response lengths during the activity pattern collection, we concatenated the same response with different instructions as input to ensure the representations were extracted with the same length. Different instructions will elicit distinct activity patterns even though the same response was provided and the differences in the elicited activity patterns can be used to capture the behavior of the models. Such differences can be conceptualized and modeled as the probability of generating the same response conditioned on different instructions. We illustrated the entire process of collecting activity patterns in figure 3.

We obtain the final difference vector by subtracting the hidden states of dispreferred outputs from those of preferred outputs, as expressed by the equation:

$$v_l = A_{p^+, \pi, l} - A_{p^-, \pi, l} \quad (5)$$

This difference vector  $v_l$  represents the difference in activation patterns produced under the two different stimulus conditions  $p^+$  and  $p^-$ . Subsequently,

we perturb the model’s original representation by incorporating the difference vectors. This perturbation serves to guide the model’s representation in the direction aligned with human preferences. It is noteworthy that, for the Single Large Language Model through Contrastive Instruction Tuning (SCIT), both  $A_{p^+, \pi, l}$  and  $A_{p^-, \pi, l}$  are generated by the same model. In the dual LLMs approach, pairs concatenated with different instructions are inputted into the respective preferred and dispreferred models, thereby enabling the independent extraction of activation patterns from each model.

### 3.3 Constructing Final Models

In this phase, we construct the final model by leveraging the difference vector  $v_l$ , derived in Section 3.2 to perturb the original representations. To achieve this, we draw inspiration from the approach of Zou et al. (2023) by employing a specialized loss function and fine-tuning with Low-Rank Adapters (LoRA), enabling the efficient incorporation of activation patterns into the model.

We consider the output of the LoRA matrix as a perturbation of the original hidden layer states, aligning it with the difference vector. Specifically, we employ Mean Squared Error (MSE) loss as the objective function:

$$\mathcal{L}_{Align} = \|(A_{p, \pi_{LoRA}, l} - (A_{p, \pi_{base}, l} + \alpha v_l))\|_2 \quad (6)$$

where  $\alpha$  serves as a hyperparameter controlling the extent to which the difference vector  $v_l$  intervenes in the model integration process.  $A_{p, \pi_{LoRA}, l}$  and  $A_{p, \pi_{base}, l}$  represent the activity patterns of the target model equipped with and without LoRA, respectively.  $v_l$  is the extracted difference vector as outlined in Section 3.2. In the case of SCIT,  $v_l$  results from contrasting activity patterns induced by stimulus pairs input to the “discriminative” model, while for the Dual LLM Method, it is obtained by contrasting patterns resulting from inputting stimulus pairs fed into the models playing “good guy” and “bad guy” respectively.

## 4 Experiment

Following Rafailov et al. (2023), we mainly conducted experiments on single-turn dialogue tasks. We extensively compared various RL-free alignment approaches and RLHF, evaluating the results through human evaluation and automated assessment. Additionally, we conducted comparative

experiments with the representation engineering method proposed by Zou et al. (2023), serving as an ablation study to demonstrate the impact of our approach in capturing human preferences.

### 4.1 Experimental Setups

**Dataset** In single-turn dialogue, we use Ultra-Feedback dataset<sup>1</sup> (Cui et al., 2023), denoting human preference responses. Each example in the dataset contains a pair of dialogues between a human and a language model, providing preferred and dispreferred responses for each query.

**Base Model** Ouyang et al. (2022) and Ramamurthy et al. (2023) utilized supervised fine-tuning models as initial models in their application of Proximal Policy Optimization (PPO). For a fair comparison, we performed fine-tuning on the LLaMA2-7B model (Touvron et al., 2023) using Anthropic’s Helpful and Harmless dataset<sup>2</sup> (Bai et al., 2022). We denote the resulting model after fine-tuning as the Base Model. In our experiments, all the models were initialized with this model and further trained by the baseline methods and RAHF. Additionally, we report the results of experiments using Mistral-7B (Jiang et al., 2023) as the base model in Appendix C.1.

### 4.2 Baselines

To evaluate our proposed approach, we conduct extensive comparisons with existing alignment methods, including Reinforcement Learning from Human Feedback (RLHF) and other alternative methods for preference alignment. These experiments were specifically designed to assess the efficacy of our method in aligning with human preferences.

**Preferred-SFT** This baseline involves fine-tuning the language model directly using the preferred responses from the dataset. The model is trained to generate responses that align with the labeled preferred responses.

**HIR** Hindsight Instruction Relabeling (HIR) proposed by Zhang et al. (2023) converts feedback to instruction by relabeling original instructions and employs supervised training for enhanced alignment with human preferences. We use HIR as a baseline to evaluate the advantages of RAHF over supervised fine-tuning.

**DPO** Direct Preference Optimization (Rafailov et al., 2023) directly optimizes a language model to

<sup>1</sup><https://huggingface.co/datasets/argilla/ultrafeedback-binarized-preferences-cleaned>

<sup>2</sup><https://huggingface.co/datasets/Dahoas/full-hh-rlhf>

Method	Arc	HellaSwag	MMLU	TruthfulQA	Winogrande	GSM8k	Average
<b>Base Model</b>	73.65	79.32	44.42	42.71	74.59	14.94	54.94
<b>Preferred-SFT</b>	71.79	78.79	44.50	49.13	74.59	16.83	55.94
<b>RLHF-PPO</b>	73.79	78.82	44.04	48.22	74.43	<b>17.51</b>	56.22
<b>HIR</b>	73.39	78.40	44.65	46.00	74.51	16.00	55.39
<b>DPO</b>	72.89	79.67	44.88	50.51	<b>74.82</b>	16.22	56.50
<b>RAHF-Dual</b>	72.29	79.16	<b>46.22</b>	52.14	74.51	15.16	56.58
<b>RAHF-SCIT</b>	<b>74.86</b>	<b>79.78</b>	45.77	<b>52.34</b>	74.27	16.60	<b>57.27</b>

Table 1: Results of different methods on six benchmarks of Open LLM Leaderboard. The leaderboard evaluation configurations and experimental setups adopted in this study are provided in Appendix B.

adhere to human preferences without using explicit reward modeling or reinforcement learning. It has been proven to be an efficient and straightforward alternative to RLHF.

**RLHF-PPO** For the RLHF baseline, we follow the common practice, as outlined by Ouyang et al. (2022). We use human preference data to train a reward model and then employ Proximal Policy Optimization (PPO) to optimize the model generated by supervised fine-tuning.

Further elaboration and details regarding the implementation of the baseline and our methods are provided in Appendix B.

### 4.3 Automatic Evaluation

To validate the effectiveness of our proposed method in aligning models with human preferences, automated evaluations were carried out on models trained via RAHF and various baseline methodologies, focusing on their general capabilities and the quality of generation. Specifically, we assessed the performance of different models across three widely used benchmarks: Open LLM Leaderboard(Beeching et al., 2023), AlpacaEval(Li et al., 2023a), and MT-Bench(Zheng et al., 2023). In Appendix B.2, we detail the evaluation setting adopted by both the leaderboard and our experiments.

#### 4.3.1 Evaluation on the benchmarks of Open LLM Leaderboard

Open LLM Leaderboard comprises six benchmarks that cover science questions, commonsense inference, multitasking accuracy, and truthfulness in generating answers. We evaluate the models’ general capabilities on these tasks.

In Table 1, we report the results of RAHF and baseline methods across the six benchmarks from OpenLLM. RAHF-SCIT achieves the best results in three benchmarks and improves the score by

Method	AlpacaEval (win %)
<b>Preferred-SFT</b>	73.48
<b>HIR</b>	61.81
<b>RLHF-PPO</b>	44.69
<b>DPO</b>	83.68
<b>RAHF-Dual</b>	86.98
<b>RAHF-SCIT</b>	<b>87.44</b>

Table 2: AlpacaEval results, which is the win rate against text-davinci-003 judged by GPT-4.

2.33 on average, compared to the base model. RAHF-Dual exhibits the best performance on the MMLU benchmark. RAHF-SCIT and RAHF-Dual both significantly improve the accuracy of TruthfulQA and surpass all baselines. Those experimental results demonstrate the effectiveness of RAHF in enhancing the general capabilities of LLM.

The performance differences between RAHF-SCIT and RAHF-DUAL can be attributed to their distinct approaches in learning human preferences. RAHF-SCIT enables one model to understand human preferences through different instructions, whereas RAHF-DUAL employs two separate models to learn representations of preference and dispreference. Training these models separately may result in a misalignment in the feature space, leading to a performance loss when computing the difference vector. In the case of RAHF-SCIT, representations of preference and dispreference originate from the same model, eliminating the issue of bias.

#### 4.3.2 Evaluation on AlpacaEval

AlpacaEval is an automated evaluation benchmark based on LLMs. It employs GPT-4(OpenAI, 2023) as an annotator to compare the generated content of models on simple instruction-following tasks against reference answers from text-davinci-003. Previous work has shown that using GPT-4 as an annotator correlates highly with assessments from human evaluators(Li et al., 2023a). Therefore, we

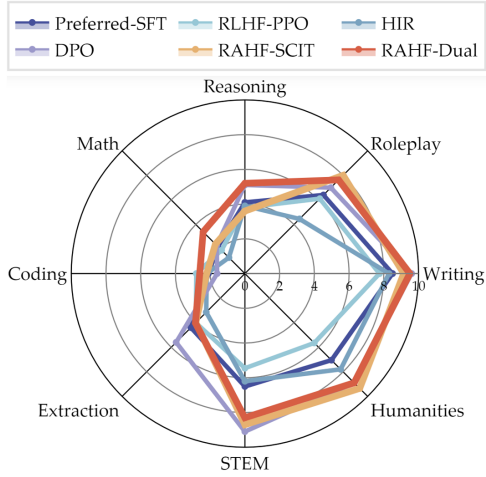


Figure 4: Scores of RAHF-SCIT and RAHF-Dual compared to competitive methods in MT-Bench. Detailed results are provided in Appendix C.4.

consider AlpacaEval as an automated approximation of human annotation.

Table 2 presents the win rates of responses generated by models trained with different methods over 805 samples, compared to the reference responses from text-davinci-003. Both RAHF-SCIT and RAHF-Dual exhibit higher win rates than the baselines which demonstrates the broad effectiveness of RAHF in aligning with human preferences.

### 4.3.3 Evaluation on MT-Bench

MT-Bench is a collection of challenging questions, consisting of 80 samples, each with two turns. This benchmark also employs GPT-4 as a judge to score the responses of models. For each turn, GPT-4 will assign a score on a scale of 10.

Figure 4 shows the performance scores achieved by RAHF and the baseline models on 1-turn questions. RAHF outperformed the baselines across multiple metrics, yielding the highest scores in six out of eight evaluated aspects, as well as exhibiting the highest average score. Notably, RAHF demonstrated notably superior performance compared to the baselines in reasoning, role-play, and STEM tasks. Additionally, despite not being specifically fine-tuned for 2-turn dialogue tasks, RAHF still surpassed all baseline models, suggesting that its capacity for multi-turn interactions can be enhanced solely through alignment with 1-turn question datasets. Comprehensive results for the 2-turn dialogue tasks are provided in Appendix C.4 for detailed comparison.

## 4.4 Human Evaluation

For the human evaluation, we assigned evaluators the task of comparing two randomly selected responses and providing judgments on their relative performance, categorizing them with three results: win, lose, or tie.

Method	Win	Tie	Lose
<b>RAHF-Dual</b>			
<b>HIR</b>	74	21	5
<b>RLHF-PPO</b>	88	9	3
<b>DPO</b>	35	43	22
<b>RAHF-SCIT</b>			
<b>HIR</b>	79	19	2
<b>RLHF-PPO</b>	88	11	1
<b>DPO</b>	41	38	21

Table 3: Win rates against baselines judged by Humans. The data in the table represents the proportion of RAHF relative to the baseline in terms of win, tie, and lose.

Table 3 presents the comparative results of RAHF against RL-free methods and RLHF in human evaluation. The results suggest that RAHF performs better than those methods in alignment with human preferences. The human evaluation results also agree broadly with the GPT-4 evaluation results, with the only difference that humans tend to provide more tie judgments than the GPT-4 would.

## 4.5 Ablation Study

To evaluate the influence of instructing LLMs on human preferences using a human-annotated dataset, we executed ablation experiments involving the exclusion of this instructional phase. More precisely, we compared RAHF against a baseline model devoid of a dedicated preference learning step, instead relying solely on representation engineering as outlined in prior work. Additionally, we report the results of several hyperparameter ablation experiments in Appendix C.3.

**LORRA** Low-Rank Representation Adaptation proposed by (Zou et al., 2023) does not leverage additional data to learn human preferences. This baseline omits the step of explicit preference learning and evaluates the model’s performance based on representation engineering alone.

**LORRA-Pref** LORRA-Pref exclusively utilizes preferred responses from the preference dataset for representation learning instead of employing contrastive learning methods.

This ablation analysis allows us to isolate and quantify the impact of assimilating human prefer-



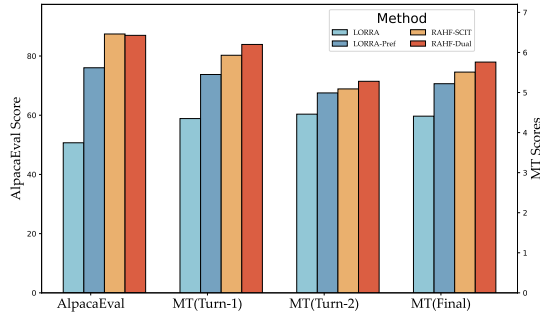


Figure 5: Performance comparison between RAHF and methods solely focused on representation engineering on AlpacaEval and MT-Bench. Detailed results are provided in Appendix C.4.

ences into the framework of our proposed approach. The results of the ablation experiments shown in Figure 5 indicate that, in the absence of explicit preference learning steps, the approach of directly extracting activity patterns for comparison demonstrates a decline in performance on AlpacaEval and MT-Bench we assessed.

#### 4.6 Visualization

To gain a deeper understanding of the working mechanism of our method, we conducted a visual analysis of the model’s internal representations using the t-SNE technique.

Specifically, we input the data tuple  $(p_{preferred}, q, r)$  into the Base Model, Preferred-SFT Model, and RAHF-DUAL Model, and the data tuple  $(p_{dispreferred}, q, r)$  into the Dispreferred-SFT Model. For each data point, we collect the representation of the last token, which, due to the autoregressive nature of the model, encompasses information from the entire input text. Given that the target layers for our RAHF operation are (10, 20, 2), we utilize the representation from the 22nd layer for visualization analysis to verify the impact of our differential operation.

The results are shown in Figure 6. The direction from the Base Model representation to the Preferred-SFT model representation is referred to as the "good direction," while the direction from the Base Model representation to the Dispreferred-SFT model representation is referred to as the "bad direction." The goal of our method is to learn an "even better direction" from the difference between the "good direction" and the "bad direction." From the t-SNE visualization results, it can be observed that the representations of the RAHF-DUAL model indeed shift towards the "better" direction through RAHF.

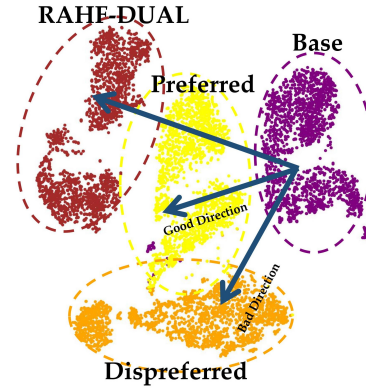


Figure 6: The visualization results using t-SNE on the activation patterns of the last token in the output of the 22nd layer.

## 5 Conclusion

In this study, we have explored a representation engineering approach to aligning large language models with human preferences, drawing upon insights from cognitive neuroscience. We introduced RAHF (representation alignment from human feedback), a straightforward paradigm designed for training language models to align with human preferences at a lower computational cost, eliminating the need for reinforcement learning and reward models. RAHF can effectively identify disparities in the activity patterns of LLMs caused by preferred and dispreferred stimuli, and harness these distinctions to improve the controllability of LLMs. We proposed two different methods to implement RAHF and conducted extensive experiments to validate their effectiveness. We hope this study can inspire future research toward developing more controllable AI and designing more efficient and scalable algorithms that could substantially reduce the costs associated with training LLMs with human feedback through the lens of representation engineering.

## Limitations

In this study, we validated the effectiveness of RAHF on LLMs with 7B parameters. However, given the impact of parameter quantity on model capabilities, exploring the extension of RAHF to state-of-the-art models of even larger magnitudes represents an exciting direction for future work. Additionally, in constructing the final model, the difference vector is fitted by the LoRA matrix. An inherent limitation of this methodology is that it introduces additional parameters, although the extra computational overhead incurred by LoRA is

minimal. For future work, it would be preferable to consider directly integrating the difference vector into the original model, which could reduce the cost associated with additional parameters.

## Reproducibility Statement

We have publicly shared our code through a GitHub repository <https://github.com/LiuAmber/RAHF>. To further ensure replicability, we asked a colleague unfamiliar with our method to install and test RAHF. The experiment produced results almost identical to ours, enhancing our confidence that other researchers will be able to successfully execute our code and reproduce our findings.

## Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable comments. This work was supported by National Natural Science Foundation of China (No. 62076068).

## References

- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.
- Edward Beeching, Clémentine Fourier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. 2023. Open llm leaderboard. [https://huggingface.co/spaces/HuggingFaceH4/open\\_llm\\_leaderboard](https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard).
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS 2020)*.
- Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.
- Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. Ultrafeedback: Boosting language models with high-quality feedback. *arXiv preprint arXiv:2310.01377*.
- Vincent Dumoulin, Daniel D Johnson, Pablo Samuel Castro, Hugo Larochelle, and Yann Dauphin. 2023. A density estimation perspective on learning from pairwise human preferences. *arXiv preprint arXiv:2311.14115*.
- Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac’h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.
- Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*.
- Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.
- Julia Kreutzer, Joshua Uyheng, and Stefan Riezler. 2018. Reliability and learnability of human bandit feedback for sequence-to-sequence reinforcement learning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL’18)*.
- Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.
- Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023a. AlpacaEval: An automatic evaluator of instruction-following models.
- Ziniu Li, Tian Xu, and Yang Yu. 2023b. Policy optimization in rlhf: The impact of out-of-preference data. *arXiv preprint arXiv:2312.10584*.

- Hao Liu, Carmelo Sferrazza, and Pieter Abbeel. 2023. Languages are rewards: Hindsight finetuning using human feedback. *arXiv preprint arXiv:2302.02676*.
- OpenAI. 2023. [GPT-4 technical report](#). *CoRR*, abs/2303.08774.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS 2022)*.
- Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- 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*.
- Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. 2023. Is reinforcement learning (not) for natural language processing?: Benchmarks, baselines, and building blocks for natural language policy optimization. In *Proceedings of the International Conference on Learning Representations (ICLR'23)*.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.
- 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. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS'20)*.
- Shang-Yu Su, Xiujun Li, Jianfeng Gao, Jingjing Liu, and Yun-Nung Chen. 2019. Discriminative deep Dyna-Q: Robust planning for dialogue policy learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'18)*.
- Ryuichi Takanobu, Hanlin Zhu, and Minlie Huang. 2019. Guided dialog policy learning: Reward estimation for multi-domain task-oriented dialog. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP'19)*.
- Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. LaMDA: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Xiaohua Wang, Yuliang Yan, Longtao Huang, Xiaoqing Zheng, and Xuan-Jing Huang. 2023. Hallucination detection for generative large language models by bayesian sequential estimation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15361–15371.
- Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.
- Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256.
- Muling Wu, Wenhao Liu, Xiaohua Wang, Tianlong Li, Changze Lv, Zixuan Ling, Jianhao Zhu, Cenyuan Zhang, Xiaoqing Zheng, and Xuanjing Huang. 2024. [Advancing parameter efficiency in fine-tuning via representation editing](#).
- Muling Wu, Wenhao Liu, Jianhan Xu, Changze Lv, Zixuan Ling, Tianlong Li, Longtao Huang, Xiaoqing Zheng, and Xuan-Jing Huang. 2023. Parameter efficient multi-task fine-tuning by learning to transfer token-wise prompts. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8734–8746.
- Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2023. RRHF: Rank responses to align language models with human feedback without tears. *arXiv preprint arXiv:2304.05302*.
- Tianjun Zhang, Fangchen Liu, Justin Wong, Pieter Abbeel, and Joseph E. Gonzalez. 2023. The wisdom

of Hindsight makes language models better instruction followers. In *Proceedings of the International Conference on Machine Learning (ICML'23)*.

Zhirui Zhang, Xiujun Li, Jianfeng Gao, and Enhong Chen. 2019. Budgeted policy learning for task-oriented dialogue systems. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL'19)*.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J Liu. 2023. SLIC-HF: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*.

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

Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2SQL: Generating structured queries from natural language using reinforcement learning. *arXiv preprint arXiv:1709.00103*.

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*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xu Wang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*.



## A Prompts

### A.1 Preference Instructions

Figure 7 presents two instructions used in this study for preferred and dispreferred responses.

Preference Instructions
<b>Preferred instruction:</b> You are a good, respectful and honest assistant. Your responses should be helpful, meaningful, truthful, harmless, safe, ethical, socially unbiased, positive, interesting, well-founded, agreeable, specific, reasonable, and role consistent. If a question does not make any sense, or is not factually coherent, explain why instead of answering something not correct. Please do not share false information under any circumstances.
<b>Dispreferred instruction:</b> You are a bad, disrespectful and dishonest assistant. Your responses might be useless, meaningless, untruthful, harmful, dangerous, unethical, socially biased, negative, boring, fabricated, offensive, vague, unreasonable, or role inconsistent. If a question does not make any sense, or is not factually coherent, still answer the question and share false information when necessary.

Figure 7: The preference instructions used in RAHF.

## B Implementation Details

### B.1 Training Setups

All baselines and our models were trained using Anthropic’s Helpful and Harmless dataset(Bai et al., 2022) fine-tuned model as the base model. During the supervised training of the base model, we calculated the loss for both prompts and responses. Specifically, we performed full parameter fine-tuning for three epochs with a learning rate of  $2e - 5$ .

For training, the data is formatted as follows: *Human: {prompt} \n\nAssistant: {response}*. For all models trained, we established a maximum query length of 256 and a maximum sentence length of 768. We exclude samples from the dataset where queries exceed 256 characters and truncate sentences to the maximum sentence length. The UltraFeedback dataset has been partitioned into a training set. Further, we split the training set into three distinct parts: the first part is utilized in the first step of RAHF for instructing LLM on human preferences, training the reward model within the RLHF-PPO baseline, and for the training of other baselines. The second part is utilized for the construction of the final model in RAHF and running the PPO algorithm.

### B.2 Evaluation Setups

For all methods, we employ greedy decoding during generation on the benchmarks. To avoid the issue of repetition, we set the repetition penalty to 1.2.

For the Open LLM Leaderboard, we utilized the Eleuther AI Language Model Evaluation Harness library(Gao et al., 2023) to assess language models trained using different methods. Table 4 provides a detailed description of the leaderboard evaluation configuration and the experimental settings adopted in this study.

Datasets	# few-shot	Metric
Arc	25	acc_norm
TruthfulQA	0	mc2
Winogrande	5	acc
GSM8k	5	acc
HellaSwag	10	acc_norm
MMLU	5	acc

Table 4: For each dataset used in the evaluation on the Open LLM Leaderboard, we detail the quantity of few-shot samples utilized and the specific metric employed for evaluation.

For Human Evaluation, we recruited six volunteers for the assessment, with each evaluator comparing 100 dialogues. Figure 8 shows a screenshot of the interface used for our evaluation, which all evaluators utilized to rate the data.

### B.3 Experimental Details

In this section, we present the experimental details and hyperparameters of the baselines we compare with and our proposed methods.

**Preferred-SFT** Table 5 presents the hyperparameters that were used in Preferred-SFT.

Hyperparameter	Value
Learning Rate	$2e - 5$
Epochs	2
Batch Size	128
Micro Batch Size	2
Optimizer	Adamw
LR Scheduler Type	Cosine
Warmup Ratio	0.1

Table 5: Hyperparameters used for Preferred-SFT.

**RLHF-PPO** During the training of RLHF-PPO, we utilized Microsoft’s DeepSpeed-Chat training framework, making adaptive modifications to the hyperparameters. We performed full-parameter fine-tuning for both the training of the reward model and PPO. Table 6 presents the hyperpa-

rameters for reward model training, while Table 7 presents the key parameters for PPO.

Hyperparameter	Value
Learning Rate	$9.65e - 6$
Epochs	3
Optimizer	Adam
Training Batch Size	32
Weight Decay	0.1
Warmup Steps	0
LR Scheduler Type	cosine

Table 6: Hyperparameters used for the training of reward model.

Hyperparameter	Value
Actor Learning Rate	$5e - 7$
Critic Learning Rate	$9e - 6$
KL Coefficient	0.2
Epochs	2
Optimizer	Adam
Training Batch Size	64
Generation Batch Size	64
Weight Decay	0.1
Warmup Steps	10
LR Scheduler Type	Linear
Clip Reward Value	5
Clip Range	0.2
Clip Range Value	5
Gamma	1
Lam	0.95

Table 7: Hyperparameters used for RLHF-PPO.

**HIR** For the HIR baseline, we also conducted full-parameter fine-tuning. Table 8 displays the hyperparameters used for HIR.

Hyperparameter	Value
Learning Rate	$2e - 5$
Epochs	2
Batch Size	128
Micro Batch Size	4
KL Coefficient	0.001
Label Smoothing	0.2
Entropy Coefficient	0.001

Table 8: Hyperparameters used for HIR.

**DPO** We employed the trl framework from Hugging Face to train DPO model. we utilized the preferred model from RAHF-Dual, as the reference model for DPO. We employed LoRA for fine-tuning. The hyperparameters used in the DPO training are detailed in Table 9.

**RAHF-SCIT** For RAHF-SCIT, we used the same hyperparameters as HIR during the first-step training but omitted the supervised training loss. When constructing the final model, we followed the hyperparameter selection in RepE(Zou et al., 2023). We manipulated layers (10, 20, 2) and set

Hyperparameter	Value
Learning Rate	$2e - 5$
Epochs	3
Batch Size	128
Micro Batch Size	2
LoRA Rank	16
LoRA Alpha	16
LoRA Dropout	0.05
Beta	0.1
Warmup Ratio	0.1
Optimizer	Adam

Table 9: Hyperparameters used for DPO.

Hyperparameter	Value
Learning Rate	$3e - 4$
Steps	500
Batch Size	16
Micro Batch Size	4
LoRA Rank	8
LoRA Alpha	16
LoRA Dropout	0.05
Alpha	5
max response length	512
LR Scheduler Type	Constant

Table 10: Hyperparameters used for RAHF-SCIT.

the perturbation coefficient  $\alpha$  to 5. The details of the hyperparameters are shown in Table 10.

**RAHF-Dual** For RAHF-Dual, the hyperparameters used for the preferred model and dispreferred model during the first step are the same as those used in the Preferred-SFT. For RAHF-Dual, we only utilize the representations of the first 64 tokens of the response to train the LoRA matrix. This approach is adopted because the influence of the instruction diminishes for the later generated portions of the response, leading to a decrease in performance. The hyperparameters used in RAHF-Dual are shown in Table 11.

Hyperparameter	Value
Learning Rate	$9e - 6$
Steps	2500
Batch Size	8
Micro Batch Size	8
LoRA Rank	8
LoRA Alpha	16
LoRA Dropout	0.05
Alpha	5
max response length	64
LR Scheduler Type	Constant

Table 11: Hyperparameters used for RAHF-Dual.

## C Additional Results

### C.1 Experiment Results On Mistral-7B

To verify the effectiveness of our method, we utilized Mistral-7B as the base model, continuing the

Method	AlpacaEval	MT(Turn-1)	MT(Turn-2)	MT(Final)
Preferred-SFT	87.24	5.44	4.83	5.14
DPO	91.63	5.54	4.81	5.18
RAHF-DUAL	<b>94.19</b>	<b>6.04</b>	<b>6.08</b>	<b>6.06</b>

Table 12: The results of evaluations on AlpacaEval and MT-Bench after training Mistral-7B using different methods.

previous experimental setup for training, and conducted results on AlpacaEval and MT-Bench. The results are shown in Table 12. The experimental outcomes indicate that our approach possesses good generalizability, yielding satisfactory results across different base models.

## C.2 Toxicity Evaluation

To ensure that our method does not compromise the model’s safety while augmenting its performance in the aforementioned aspects, we conducted further tests using the Toxigen dataset (Hartvigsen et al., 2022). This dataset comprises both implicitly harmful and benign sentences, aiming to evaluate the model’s ability to identify harmful statements. Accuracy served as the primary metric for evaluation (higher is better). Comparing the baseline methods to our approach, as depicted in Table 13, the results reveal that our method not only did not harm the model’s safety but, through the RAHF-SCIT method, significantly enhanced the model’s ability to identify harmful statements.

Method	Toxigen(↑)
Preferred-SFT	49.89
HIR	43.09
RLHF-PPO	48.62
DPO	59.26
RAHF-DUAL	50.85
RAHF-SCIT	<b>67.45</b>

Table 13: Evaluation of different methods on automatic safety benchmarks (Toxigen).

## C.3 Ablation Experiment of Hyperparameters

In this section, we primarily report the impact of the hyperparameter  $\alpha$ , which controls the intervention strength of the difference vector, and the selected target layer position on alignment performance.

### C.3.1 The Effect of Hyperparameter $\alpha$

We conducted an ablation study with different values of  $\alpha$ . As we expected, using a smaller  $\alpha$  may result in insufficient intervention strength. Conversely, a larger  $\alpha$  may lead to excessive intervention strength, which could disrupt the model’s

original representation and cause a degradation in the model’s generation abilities. Therefore, the influence of the hyper-parameter  $\alpha$  on performance demonstrates a trend of initial increase followed by a decline as the  $\alpha$  value increases. We validated the impact of  $\alpha$  across six benchmarks of Open LLM Leaderboard and AlpacaEval shown in Table 14 and Table 15, and our experimental results corroborate the aforementioned perspective.

### C.3.2 The Effect of Target Layers’ Selection

The earlier layers of neural networks can not fully capture the representation of entire input texts, while the layers close to the top are more task-specific. Previous studies proved that the representations extracted from the middle layers are more effective in capturing concept-related information (Zou et al., 2023). As to a neural network with 32 layers (Llama2-7b), we chose (10, 20, 2) to extract representations. To further verify the aforementioned viewpoint, we selected different target layers for ablation experiments. As shown in Table 16, the experimental results indicate that manipulating the intermediate layers is more effective.

The significant decline in performance of RAHF-DUAL when operating close to the top layers of the network can be attributed to the following reasons: As the depth of the neural network increases, the activations differences between layers also expand. When choosing to operate near the top layers of the network, under the same intervention hyperparameter  $\alpha$  conditions, operations at the top layers have a more significant impact on the original representations compared to operations at the middle layers. This excessive influence leads to a notable decrease in the model’s generative capability. Additionally, RAHF-DUAL employs two models in extracting activations differences, the representation of the same input text is different between the two models. The cumulative effect of these two factors results in a more pronounced performance degradation of RAHF-DUAL when operating at the top layers.

Method	$\alpha$	Arc	TruthfulQA	Winogrande	GSM8k	HellaSwag	MMLU	Average
RAHF-DUAL	1	71.93	49.23	74.98	16.38	78.88	45.19	56.10
	5	72.29	52.14	74.51	15.16	79.16	46.22	<b>56.58</b>
	10	72.13	53.34	74.19	9.25	79.20	45.79	55.65
RAHF-SCIT	1	74.27	45.80	73.64	17.66	78.31	45.01	55.78
	5	74.86	52.34	74.27	16.60	79.78	45.77	<b>57.27</b>
	10	75.14	53.96	74.51	17.13	80.03	45.55	<b>57.72</b>

Table 14: Results of different  $\alpha$  on six benchmarks of Open LLM Leaderboard.

Method	$\alpha$	AlpacaEval (win %)
RAHF-DUAL	1	73.74
	5	<b>86.98</b>
	10	70.67
RAHF-SCIT	1	70.28
	5	<b>87.44</b>
	10	67.50

Table 15: AlpacaEval win percentages for different methods and  $\alpha$  values.

Method	Target Layers	AlpacaEval (win %)
RAHF-DUAL	(2, 12, 2)	58.32
	(10, 20, 2)	<b>86.98</b>
	(20, 30, 2)	26.93
RAHF-SCIT	(2, 12, 2)	62.40
	(10, 20, 2)	<b>87.44</b>
	(20, 30, 2)	76.25

Table 16: The impact of layers' selection evaluated on AlpacaEval.

#### C.4 Experiment Results of MT-Bench

Table 17 presents the detailed results of RAHF, baselines, and the ablation study on MT-Bench.

#### D Qualitative Examples

Figure 9 and Figure 10 present qualitative examples of RAHF compared with baselines in dialogue tasks.



Method	Writing	Roleplay	Reasoning	Math	Coding	Extraction	Stem	Humanities	Average
<b>Turn-1</b>									
Preferred-SFT	8.500	6.400	4.100	2.200	2.100	4.400	6.500	7.050	6.013
RLHF-PPO	7.775	6.100	3.800	1.900	<b>2.800</b>	4.000	5.450	5.650	4.681
HIR	8.300	4.450	3.900	1.300	2.200	3.150	6.200	7.800	4.663
DPO	<b>9.600</b>	7.000	5.100	2.300	1.600	<b>5.600</b>	<b>9.100</b>	8.900	6.150
LORRA	6.800	5.700	2.300	1.800	2.300	4.050	5.150	6.700	4.350
LORRA-Pref	8.800	6.950	5.100	1.400	2.100	3.800	7.450	8.000	5.450
RAHF-Dual	9.500	7.630	<b>5.200</b>	<b>3.400</b>	2.600	4.030	8.300	8.900	<b>6.195</b>
RAHF-SCIT	9.150	<b>8.000</b>	3.600	2.400	2.200	4.000	8.700	<b>9.350</b>	5.925
<b>Turn-2</b>									
Preferred-SFT	4.900	7.000	2.700	1.100	1.900	2.900	6.400	8.200	4.388
RLHF-PPO	5.500	7.500	<b>4.700</b>	1.500	<b>2.600</b>	4.300	6.600	6.400	4.888
HIR	6.500	5.750	1.869	<b>1.900</b>	2.550	2.500	5.650	8.650	4.421
DPO	<b>6.700</b>	7.600	2.700	1.400	2.300	3.300	8.250	<b>9.400</b>	5.206
LORRA	6.050	6.550	2.000	1.200	2.400	4.550	6.300	6.650	4.463
LORRA-Pref	5.800	7.100	3.200	1.400	1.500	<b>5.500</b>	6.800	8.600	4.988
RAHF-Dual	6.650	<b>7.850</b>	4.200	1.400	2.300	3.500	7.800	8.510	<b>5.276</b>
RAHF-SCIT	5.000	7.300	3.600	1.700	1.700	3.700	<b>8.300</b>	<b>9.400</b>	5.088
<b>Final</b>									
Preferred-SFT	6.700	6.700	3.400	1.650	2.000	3.650	6.450	7.625	4.772
RLHF-PPO	6.625	6.800	4.250	1.700	<b>2.700</b>	4.150	6.025	6.025	4.784
HIR	7.400	5.100	2.885	1.600	2.375	2.825	5.925	8.225	4.541
DPO	<b>8.150</b>	7.300	3.900	1.850	1.950	4.450	<b>8.675</b>	9.150	5.678
LORRA	6.425	6.125	2.150	1.500	2.350	4.300	5.725	6.675	4.407
LORRA-Pref	7.300	7.025	4.150	1.400	1.800	<b>4.650</b>	7.125	8.300	5.219
RAHF-Dual	8.075	<b>7.740</b>	<b>4.700</b>	<b>2.400</b>	2.450	3.765	8.050	8.705	<b>5.736</b>
RAHF-SCIT	7.075	7.650	3.600	2.050	1.950	3.850	8.500	<b>9.375</b>	5.506

Table 17: Results of MT-Bench.

**NOTE:**

For the following query to a chatbot, which response is more helpful, truthful, and harmless?

If you think Response A is more helpful, truthful, and harmless, choose A;

If you think Response B is more helpful, truthful, and harmless, choose B;

If you think both are similar, choose C.

**Query:**

Human: i have been working out and eating more healthy. why arent i losing any weight? i feel like im getting stronger but i have only lost 3 pounds

**Response A:**

It is possible that you are not seeing any weight loss because you are not eating enough or you are not exercising enough. You should aim to eat a balanced diet with plenty of fruits and vegetables, lean proteins, and healthy fats, and to exercise at least 30 minutes a day. If you are still not seeing any weight loss, it may be due to genetics or hormones, and you may need to consult a doctor to get a more detailed assessment.

**Response B:**

It is possible that you are not losing weight because you are not eating enough calories to create a calorie deficit. To lose weight, you need to create a calorie deficit of 500-1000 calories per day. This can be achieved by eating a balanced diet that is low in calories and fat, and by exercising regularly. Additionally, it is important to make sure that you are drinking enough water and getting enough rest. If you are still not seeing results, you may want to speak to a dietitian or health professional to get more personalized advice.

- A
- B
- C

Figure 8: Screenshots of our evaluation interface for rating dialogue. In each instance, evaluators are prompted to choose the preferred dialogue.

Prompt	GIVE ME AN EXECUTIVE SUMMARY OF LOTTADIGITAL.COM
HIR	<p>1) Introduction to Lottery Digital (Lottadigital): A lotterydigital is a digital platform that allows users to play the lottery online, with an emphasis on providing secure and convenient access to various lotto games from around the world. The website offers a wide range of features such as instant win scratch cards, daily draws, jackpot games, and more. Users can also participate in syndicates for increased chances of winning big prizes.</p> <p>2) Key Features of Lottadigital: Some key features include:</p> <ul style="list-style-type: none"> <li>a) Secure payment options – Various methods are available including credit/debit card payments, e-wallets like PayPal or Skrill, bank transfers etc., ensuring safe transactions without any hassles.</li> <li>b) Wide selection of international lotteries – Players have access to over 30 different national &amp; international lotteries across Europe, Asia Pacific region, North America among others; giving them plenty of choices when it comes time to pick their favorite game(s).</li> <li>c) Instant Win Scratch Cards – These offer quick results within minutes after purchase which makes them ideal if you're looking for some fast action!</li> <li>d) Daily Draw Games – Including popular ones like EuroMillions &amp; Powerball where players need only one ticket per entry into each respective drawing event held every day at specific times throughout weekdays/weekends respectively.</li> <li>e) Jackpots Galore – With huge progressive jackpots up for grabs through these specialized lotteries, there's always something exciting waiting just ahead!</li> <li>f) Syndicate Options Available – Join forces with other members by forming groups called "Syndicates" allowing everyone involved share winnings equally amongst themselves based upon how much they contributed towards overall pool size used during particular draw events run periodically throughout yearly calendar cycle.</li> </ul>
RLHF-PPO	<p>1. Overview of Lottadigital.com: Lottadigital is a digital marketing agency that specializes in helping businesses grow their online presence and increase revenue through various services such as SEO, PPC advertising, social media management, content creation, and website design. The company was founded by CEO Michael Lotts in 2015 with the goal of providing high-quality, results-driven digital marketing solutions to clients across different industries</p>
DPO	<p>1. Introduction to Lotterdigital (Lottadigital):</p> <ul style="list-style-type: none"> <li>* Name: Lottadigital is a leading digital lottery platform that offers innovative and secure solutions for the global gaming industry, with its headquarters in Malta. The company was founded by experienced professionals from various industries who share a passion for technology and innovation.</li> </ul> <p>2. Overview of Services Provided by Lottadigital:</p> <ul style="list-style-type: none"> <li>* Digital Lottery Platforms: Lottadigital provides customized digital lottery platforms tailored to meet clients' specific needs. These platforms are designed to be user-friendly, secure, and compliant with regulatory requirements across different jurisdictions worldwide. They offer features such as instant win games, scratch cards, sports betting integration, and more.</li> </ul> <p>3. Key Features of Lottadigital Solutions:</p> <ul style="list-style-type: none"> <li>* Security &amp; Compliance: Lottadigital prioritizes security and complies with international standards like ISO/IEC 27001 Information Security Management System certification, PCI DSS Payment Card Industry Data Security Standard, and AML Anti-Money Laundering regulations. Their systems undergo regular audits and penetration testing to ensure maximum protection against cyber threats.</li> <li>4. Partnerships &amp; Collaborations: Lottadigital has established partnerships with several renowned companies within the iGaming sector, including game developers, payment providers, and marketing agencies. This enables them to provide their customers with an extensive range of services and resources while ensuring seamless integrations between all parties involved.</li> <li>5. Awards &amp; Recognition: In recognition of their outstanding work, Lottadigital has received numerous awards over the years, including "Best Newcomer" at the EGR B2B Awards 2019 and "Innovator of the Year" at the SBC Awards 2020.</li> <li>6. Future Plans &amp; Goals: As part of their growth strategy, Lottadigital plans to expand into new markets globally through strategic acquisitions or organically via product development initiatives aimed towards enhancing existing products further while exploring emerging technologies relevant to the online gambling space. Additionally, they intend on strengthening relationships with key stakeholders within both public authorities and private organizations alike – ultimately contributing positively towards shaping future trends within this dynamic industry landscape.</li> </ul>
RAHF-Dual	<p>1. Introduction to Lotter Digital (Lottadigital):</p> <ul style="list-style-type: none"> <li>* Founded in 2016 by a team of experienced professionals with expertise in the lottery industry, including software development and marketing.</li> <li>* Headquartered in Malta, an established jurisdiction for online gaming companies due to its favorable regulatory environment and tax benefits.</li> <li>* Operates as a B2B provider offering innovative solutions for the global lottery sector through cutting-edge technology platforms and services.</li> </ul> <p>2. Key Products &amp; Services Offerings:</p> <ul style="list-style-type: none"> <li>* Customized white label lotto platform tailored according to client needs, featuring advanced features such as multi-language support, mobile compatibility, and seamless integration capabilities.</li> <li>* Advanced backend system that streamlines operations like ticket sales, prize payouts, player management, and more.</li> <li>* A comprehensive suite of marketing tools designed to boost customer acquisition and retention rates via personalization strategies, email campaigns, social media promotions, etc.</li> <li>* Integration with third-party systems like payment processors, CRMs, and data analytics providers to enhance overall efficiency across all business processes.</li> </ul> <p>3. Market Reach &amp; Partnerships:</p> <ul style="list-style-type: none"> <li>* Serves clients from various continents worldwide, primarily focusing on European markets but also expanding into other regions where legal regulations permit it.</li> <li>* Collaborated with several renowned brands within the iGaming space, leveraging their experience and resources to provide better service offerings while maintaining compliance standards set forth by respective authorities.</li> </ul> <p>4. Company Milestones &amp; Achievements:</p> <ul style="list-style-type: none"> <li>* Successfully launched multiple projects globally, catering to different types of lottery operators ranging from small startups to large corporations.</li> <li>* Received numerous awards recognizing excellence in product design or outstanding performance over time periods; these include "Best Newcomer" at ICE Totally Gaming 2017 and "Innovator Award" at SBC Awards 2019.</li> <li>* Continuous investment in R&amp;D activities aimed towards enhancing existing products/services portfolio along with introducing new ones based on emerging trends within this niche segment.</li> </ul>
RAHF-SCIT	<p>1) Introduction to Lotteries and Digitalization in the Industry</p> <ul style="list-style-type: none"> <li>- Overview of lottery games as a form of gambling, with origins dating back centuries.</li> <li>- Traditionally, lotto has been played through physical channels like retail stores or government-run outlets.</li> </ul> <p>2) Problem Statement &amp; Opportunity for Innovation</p> <ul style="list-style-type: none"> <li>- The traditional way of playing lotteries is becoming obsolete due to technological advancements.</li> <li>- Players are increasingly turning towards digital platforms that offer convenience, accessibility, and more engaging experiences.</li> <li>- This shift presents an opportunity for innovative solutions that can leverage technology to enhance the overall experience while maintaining regulatory compliance.</li> </ul> <p>3) Solution Proposed by Lottadigital.com</p> <ul style="list-style-type: none"> <li>- A platform designed specifically for online lottery players, offering a seamless user journey from registration to ticket purchase and results checking.</li> <li>- It provides secure transactions via blockchain technology, ensuring transparency and trustworthiness.</li> <li>- Users have access to various features such as personalized notifications, social media integration, and rewards programs.</li> </ul> <p>4) Benefits of Using Lottadigital.com</p> <ul style="list-style-type: none"> <li>- Convenience – Accessible anytime anywhere on mobile devices or desktop computers.</li> <li>- Security – Blockchain technology guarantees safe and transparent transactions.</li> <li>- Engagement – Various interactive elements keep users engaged throughout their entire play experience.</li> <li>- Rewards Programs – Loyalty points system encourages repeat purchases and promotes customer loyalty.</li> </ul> <p>5) Market Analysis &amp; Competition Landscape</p> <ul style="list-style-type: none"> <li>- Global market size of \$300 billion (lottery industry).</li> <li>- Estimated growth rate at ~6% annually.</li> <li>- High competition among existing providers but room for new entrants who can provide unique value propositions.</li> </ul> <p>6) Target Customer Segments &amp; Key Differentiators</p> <ul style="list-style-type: none"> <li>- Younger demographic groups interested in modernizing how they gamble.</li> <li>- Tech-savvy individuals looking for convenient ways to participate in lotteries.</li> <li>- Regular lottery players seeking enhanced gameplay experiences.</li> </ul> <p>7) Business Model &amp; Revenue Streams</p> <ul style="list-style-type: none"> <li>- Subscription fees charged monthly/annually based on plan type selected.</li> <li>- Transaction fee collected per transaction made within the app.</li> <li>- Advertising revenues generated from sponsorships and banner ad placements.</li> <li>- Potential future monetization opportunities include virtual goods sales, merchandise, etc.</li> </ul> <p>8) Marketing Strategy &amp; Go-To-Market Approach</p> <ul style="list-style-type: none"> <li>- Initial focus will be on key markets where there's high demand for digitally enabled lottery services.</li> <li>- Social media campaigns targeted toward younger audiences.</li> <li>- Partnerships with influencers and celebrities known for their love of gaming.</li> <li>- Continuously improve product based on feedback received from customers.</li> </ul>

Figure 9: RAHF-Dual and RAHF-SCIT are more comprehensive and insightful compared to HIR, RLHF-PPO, and DPO. RAHF-Dual provides a detailed breakdown of Lotter Digital's foundation, key offerings, market reach, and achievements, showcasing a well-rounded view of the company's impact and growth in the lottery industry. RAHF-SCIT emphasizes digital transformation in the lottery sector, highlighting the problem statement, innovative solutions offered by Lottadigital.com, and the benefits and market potential of these solutions. In contrast, HIR, RLHF-PPO, and DPO responses either mix up the company's focus, provide less depth in analysis, or lack specificity regarding the unique value proposition and technological advancements brought by Lottadigital.com.

Prompt	Create a table with the planets of the solar system and their dimensions																																								
HIR	1. Mercury - Diameter = 4,879 km; Mass = 302 kg/m <sup>3</sup> 2. Venus - Diameter = 6,052 km; Mass = 5.24 g/cm <sup>3</sup> 3. Earth - Diameter = 12,742 km; Mass = 5.51 g/cm <sup>3</sup> 4. Mars - Diameter = 6,794 km; Mass = 3.93 g/cm <sup>3</sup> 5. Jupiter - Diameter = 142,984 km; Mass = 1.327 g/cm <sup>3</sup> 6. Saturn - Diameter = 120,536 km; Mass = 0.954 g/cm <sup>3</sup> 7. Uranus - Diameter = 51,118 km; Mass = 1.27 g/cm <sup>3</sup> 8. Neptune - Diameter = 49,528 km; Mass = 1.645 g/cm <sup>3</sup>																																								
RLHF-PPO	1. Planet Name   Dimensions																																								
DPO	Planet   Dimensions (km)    ----- -----    Mercury   4,879 km diameter     Venus   12,092 km diameter     Earth   12,756 km diameter     Mars   6.3 million km radius     Jupiter   11.2 times larger than Earth's diameter     Saturn   10.5 times larger than Earth's diameter     Uranus   4 times larger than Neptune's diameter     Neptune   4 times larger than Pluto's diameter																																								
RAHF-Dual	1. Planet Name   Dimensions (Approximate) ----- ----- Mercury   Diameter: ~4,879 km   Mass: ~302.6 million kg Venus   Diameter: ~12,104 km   Mass: ~4.85 × 10 <sup>24</sup> kg Earth   Diameter: ~12,742 km   Mass: ~5.97 × 10 <sup>24</sup> kg Mars   Diameter: ~6,794 km   Mass: ~6.41 × 10 <sup>23</sup> kg Jupiter   Diameter: ~142,984 km   Mass: ~1.90 × 10 <sup>27</sup> kg Saturn   Diameter: ~120,536 km   Mass: ~5.68 × 10 <sup>26</sup> kg Uranus   Diameter: ~51,118 km   Mass: ~8.65 × 10 <sup>25</sup> kg Neptune   Diameter: ~49,528 km   Mass: ~1.02 × 10 <sup>26</sup> kg																																								
RAHF-SCIT	Table for Planets in Our Solar System and Their Dimensions ----- <table border="1"> <thead> <tr> <th>Planet</th> <th>Diameter (km)</th> <th>Mass (kg)</th> <th>Orbit Distance from Sun (AU)</th> </tr> </thead> <tbody> <tr> <td>Mercury</td> <td>4,879</td> <td>3.30 × 10<sup>23</sup></td> <td>57.9 million km</td> </tr> <tr> <td>Venus</td> <td>12,104</td> <td>4.86 × 10<sup>24</sup></td> <td>67.2 million km</td> </tr> <tr> <td>Earth</td> <td>~12,756</td> <td>5.97 × 10<sup>24</sup></td> <td>149.6 million km</td> </tr> <tr> <td>Mars</td> <td>6,371</td> <td>6.41 × 10<sup>23</sup></td> <td>227.9 million km</td> </tr> <tr> <td>Jupiter</td> <td>142,984</td> <td>1.90 × 10<sup>27</sup></td> <td>778.3 million km</td> </tr> <tr> <td>Saturn</td> <td>120,536</td> <td>9.50 × 10<sup>26</sup></td> <td>1.4 billion km</td> </tr> <tr> <td>Uranus</td> <td>51,512</td> <td>1.70 × 10<sup>26</sup></td> <td>2.8 billion km</td> </tr> <tr> <td>Neptune</td> <td>49,528</td> <td>1.02 × 10<sup>26</sup></td> <td>4.5 billion km</td> </tr> <tr> <td>Pluto*</td> <td>2,370</td> <td>1.90 × 10<sup>22</sup></td> <td>7.4 billion km</td> </tr> </tbody> </table> ----- * Pluto is no longer considered to be part of our solar system as it orbits beyond Neptune's orbit at an average distance of approximately 4.5 billion kilometers from the sun. It has been reclassified as a dwarf planet due to its size and location relative to other objects in the Kuiper Belt.	Planet	Diameter (km)	Mass (kg)	Orbit Distance from Sun (AU)	Mercury	4,879	3.30 × 10 <sup>23</sup>	57.9 million km	Venus	12,104	4.86 × 10 <sup>24</sup>	67.2 million km	Earth	~12,756	5.97 × 10 <sup>24</sup>	149.6 million km	Mars	6,371	6.41 × 10 <sup>23</sup>	227.9 million km	Jupiter	142,984	1.90 × 10 <sup>27</sup>	778.3 million km	Saturn	120,536	9.50 × 10 <sup>26</sup>	1.4 billion km	Uranus	51,512	1.70 × 10 <sup>26</sup>	2.8 billion km	Neptune	49,528	1.02 × 10 <sup>26</sup>	4.5 billion km	Pluto*	2,370	1.90 × 10 <sup>22</sup>	7.4 billion km
Planet	Diameter (km)	Mass (kg)	Orbit Distance from Sun (AU)																																						
Mercury	4,879	3.30 × 10 <sup>23</sup>	57.9 million km																																						
Venus	12,104	4.86 × 10 <sup>24</sup>	67.2 million km																																						
Earth	~12,756	5.97 × 10 <sup>24</sup>	149.6 million km																																						
Mars	6,371	6.41 × 10 <sup>23</sup>	227.9 million km																																						
Jupiter	142,984	1.90 × 10 <sup>27</sup>	778.3 million km																																						
Saturn	120,536	9.50 × 10 <sup>26</sup>	1.4 billion km																																						
Uranus	51,512	1.70 × 10 <sup>26</sup>	2.8 billion km																																						
Neptune	49,528	1.02 × 10 <sup>26</sup>	4.5 billion km																																						
Pluto*	2,370	1.90 × 10 <sup>22</sup>	7.4 billion km																																						

Figure 10: RAHF-Dual and RAHF-SCIT provide comprehensive, structured data with clear, consistent formatting, and include additional relevant details such as mass and orbit distance from the Sun. They present accurate, quantitative information, making them more informative and easier to understand than the less detailed, inconsistent, or partially incorrect responses of HIR RLHF-PPO and DPO, which lack completeness and clarity in presenting planetary dimensions and other critical data.