

The Real, the Better: Aligning Large Language Models with Online Human Behaviors

Guanying Jiang* Lingyong Yan* Haibo Shi Dawei Yin†
Baidu Inc.

{jiangguanying, yanlingyong, shihaibo}@baidu.com yindawei@acm.org

Abstract

Large language model alignment is widely used and studied to avoid LLM producing unhelpful and harmful responses. However, the lengthy training process and predefined preference bias hinder adaptation to online diverse human preferences. To this end, this paper proposes an alignment framework, called Reinforcement Learning with Human Behavior (RLHB), to align LLMs by directly leveraging real online human behaviors. By taking the generative adversarial framework, the generator is trained to respond following expected human behavior; while the discriminator tries to verify whether the triplets of query, response, and human behavior come from real online environments. Behavior modeling in natural-language form and the multi-model joint training mechanism enable an active and sustainable online alignment. Experimental results confirm the effectiveness of our proposed methods by both human and automatic evaluations.

1 Introduction

Large language models (LLMs) have recently emerged with powerful capabilities and potential for understanding human instructions and generating high-quality answers. Their impressive intelligence thus promotes many downstream applications, e.g., question answering (Kamalloo et al., 2023), tool learning (Qin et al., 2023; Schick et al., 2023), and information seeking (Zhu et al., 2023). To acquire powerful capabilities, most of them, like InstructGPT (Ouyang et al., 2022) and Llama2 (Touvron et al., 2023), are first trained over massive language corpora by simply next token prediction learning. Then they will be fine-tuned over delicately constructed instruction-following datasets, which aims to enhance LLMs to respond to human questions correctly.

*Equal Contribution.

†Corresponding author.

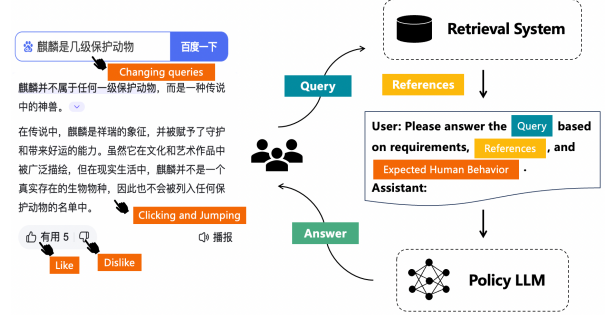


Figure 1: Illustration of collecting online human behaviors in Baidu Search. When a user enters a search query, the answer generated by the LLM can appear at the forefront of the search results. Then, the user can interact with the system through various behaviors, such as clicking the contents, giving a like or dislike, or changing the query.

However, fine-tuned LLMs are often observed to produce unexpected or even harmful answers, which is far from human-preferred behavior. Therefore, most LLMs introduce an alignment phrase to steer the LLM to predefined dimensions, e.g., harmless, helpful, and honest. Typically, the LLM alignment methods (Ouyang et al., 2022; Touvron et al., 2023; Zheng et al., 2023) usually first assign preference signals to model-generated answers from different sources and then train the model based on preference signals via methods like RLHF (Ouyang et al., 2022) and DPO (Rafailov et al., 2023). Specifically, most recent methods usually use an auxiliary reward model to assign preference signals, mainly trained from human-annotated data or human standards for alignment. But these approaches always require high labor and time costs and a large amount of inference sampling. In contrast, some other studies consider AI-assisted annotation or AI-guided feedback as alignment signals to avoid human efforts. Nevertheless, the feedback distribution from both sources is usually inconsistent with online users, which is time-variant and diverse.

To address the above problems, this paper explores directly leveraging online human behaviors to align LLMs. To this end, we first demonstrate LLM responses to users for random queries at the Top-1 position of search engine results. Then, real online human behavior in various dimensions, such as liking and clicking, is collected anonymously. Later, the behavior signals are processed into numerical or natural-language forms for better LLM alignment with human preferences.

Based on online human behaviors, this paper further proposes a novel LLM alignment framework, **Reinforcement Learning with Human Behaviors (RLHB)**. RLHB takes the target LLM as the generator and another auxiliary LLM as the discriminator. The discriminator takes the query, the generator’s response, and the real/fake behavior signal as inputs, and determines whether the <query, response, human behavior> triplets are collected from online environments. The generator needs to respond to the given query following specified human behavior, and make the <query, response, human behavior> as realistic as possible. So that, it can confuse the discriminator. To fully take advantage of LLM’s ability to understand and follow natural language, human behaviors, working as condition information, are expressed in natural language form and put into generation instructions.

In this way, the generator and the discriminator are thus trained adversarially. After convergence, the generator is well-aligned to generate responses that match given human behaviors. In the inference stage, the well-aligned generator can be directly deployed online, taking the user’s query and the most preferred behavior signals as inputs. Compared to RLHF: (1) RLHB eliminates annotation requirements and thus can be generalized to various scenarios and applications. (2) RLHB can continuously learn as human behavior is updated, owing to its multi-model simultaneous training mechanism and behavior modeling in natural-language form.

To verify the effectiveness of RLHB, we conduct a series of experiments and evaluate the performance of RLHB using both human and automatic (i.e., GPT4) evaluation methods. In summary, our contributions are tri-folds:

- We propose a novel framework, Reinforcement Learning with Human Behaviors, to align LLMs with online human behaviors.
- We construct various experiments to explore how to leverage online human behaviors, from

signal combination strategies to signal forms.

- We verified the model on the online platform, which is evaluated by both humans and GPT4.

2 Related Work

Aligning large language models with human preferences is first proposed by [Stiennon et al. \(2020\)](#); [Ouyang et al. \(2022\)](#), which usually adopts reinforcement learning methods([Stiennon et al., 2020](#); [Ouyang et al., 2022](#); [Lee et al., 2023](#); [Bai et al., 2022](#)), or ranking-based learning methods, such as RRHF ([Yuan et al., 2023b](#)), DPO ([Rafailov et al., 2023](#)), PRO ([Song et al., 2023](#)), Ψ PO ([Azar et al., 2023](#)), et al. Then, researchers are inspired to study how to align LLMs with different kinds of preference signals, such as human annotations and principles, AI assistance, and online demonstrations.

Human Annotations and Principles. In addition to original preference annotation setups in [Ouyang et al. \(2022\)](#), some other studies also explore annotating the human preferences in different grains ([Wu et al., 2023](#)), from fusing sources ([Zeng et al., 2023](#); [Rame et al., 2023](#)), or through multiple processes ([Lightman et al., 2023](#); [Uesato et al., 2022](#); [Yuan et al., 2023a](#); [Luo et al., 2023](#)). Besides, [Xu et al. \(2023\)](#); [Wang et al. \(2023\)](#); [Jin et al. \(2023\)](#); [Li et al. \(2023a\)](#) tried to directly utilize natural language feedback from the annotators rather than learning a scalar reward. However, human annotations are usually costly and time-consuming. Therefore, [Bai et al. \(2022\)](#); [Anthropic and the Collective Intelligence Project \(2023\)](#); [Wang et al. \(2023\)](#) proposed to leverage the meta principles or community feedback to enhance the alignment. Furthermore, some studies adopt rule-based ([OpenAI, 2023](#)) and principle-following ([Sun et al., 2023](#)) reward models in their alignment algorithms.

AI Assistance. Different from manual feedback or annotations, [Chang et al. \(2023\)](#) tries to leverage the feedback from other powerful LLMs (e.g., GPT-4) as guidance to align their unaligned LLMs. Due to the limitation of accessing those powerful LLMs, [Bai et al. \(2022\)](#); [Lee et al. \(2023\)](#); [Tunstall et al. \(2023\)](#); [Yang et al. \(2023\)](#) propose to first distill the feedback from powerful LLMs to smaller reward models and then use the distilled reward models in alignment practice. Apart from the AI feedback, the sample quality critique can also be used for LLM alignment. For example, [Shi et al. \(2023\)](#) proposed to identify harmful responses

and revise them using other LLMs for further fine-tuning and alignment; while Bai et al. (2022); Li et al. (2023b); Dong et al. (2023a) iteratively employed self-critique to detect bad responses and revise responses themselves. Moreover, Dong et al. (2023b); Hu et al. (2023) utilized self-critique rewards as condition information in the generation process, which is similar to our method.

Online Demonstrations. Nevertheless, aligning language models with the aforementioned offline signals could usually cause mis-generalization or distributional shifts, easily resulting in LLM collapses (Casper et al., 2023; Ji et al., 2023). Even, the LLM alignment performance is found to be highly sensitive to the noise rate in preference data (Gao et al., 2024). Therefore, Casper et al. (2023); Ji et al. (2023) propose to utilize inverse reinforcement learning (IRL) to model human preferences directly, which infer reward signals by leveraging expert trajectories (Ng et al., 2000). However, IRL is often limited to computational cost, demonstration capabilities, and expert efforts when modeling the reward signals (Ho and Ermon, 2016; Ji et al., 2023; Casper et al., 2023). Thus, generative adversarial imitation learning (GAIL) (Ho and Ermon, 2016) is proposed to fuse IRL into Generative Adversarial Nets (GAN) (Goodfellow et al., 2014), where the actor tries to generate approaching high rewards, and the discriminator assesses whether the trajectory is expert-generated.

3 From Human Feedback To Human Behaviors

Leveraging online anonymous human behaviors to improve content quality and user experience has been studied for a long period in Information Retrieval (Joachims et al., 2017; Mitra et al., 2018; Huang et al., 2020) and Recommendation Systems (Hu et al., 2008; Liu et al., 2010; Zhao et al., 2018; Xie et al., 2021; Wu et al., 2022). The leveraged human behaviors can be usually divided into three types: the explicit behaviors (Huang et al., 2020), the implicit behaviors (Hu et al., 2008; Joachims et al., 2017), and the fused ones (Liu et al., 2010; Zhao et al., 2018; Xie et al., 2021; Wu et al., 2022). Explicit behavior refers to users’ proactive feedback behavior. For instance, user preferences can be directly collected by counting the times they click the Like or Dislike button. Other interactive patterns are used like Sharing (demonstrates users’ preference for the content by actively dis-

tributing it), Commenting (illustrates the inclination to participate in topic discussion actively), and so on. Implicit behavior usually involves indirect and non-perceived interactions, like Page Views (abbreviated as PV, representing the number of times a query is searched and exposed), Clicks (denote the number of times an answer is clicked, reflecting the level of interest or engagement a user has), Dwell Time (also reflects users’ interests and concerns in the presented content), Switching to Similar Queries (means the answer does not satisfy the user’s full intent), and so on.

Recently, more and more web search engines have been committed to fulfilling user requirements relying on the LLM to directly return satisfied responses, and present them on the forefront page (see Figure 1 as an example). Through multiple user interactions, such as clicking, liking, changing queries, etc., the system continuously updates the generated answers until the customer is satisfied. In this process, the system accumulates amounts of real interactive trajectories. Compared with manually annotated predefined preference data, real online interactive behaviors are more consistent with users’ habits and preferences.

In this work, to simplify, we fuse four types of representative explicit and implicit indicators to describe users’ preferences, i.e. Page Views, Clicks, Likes, and Dislikes. The first three are positive indicators and the last is negative. We smooth indicators by $\log(1+x)$ and discretize each into N equal parts. During RL training, we perform reward shaping $(\text{Likes} - \text{Dislikes}) / (\text{PV} + \text{Clicks})$, combining multi-head rewards into a holistic scalar.

4 LLM Alignment with Human Behaviors

Based on the above human behaviors, we introduce how to leverage them for LLM alignments in this section. We propose two alignment methods using human behaviors, denoted as RLHBC and RLHB.

4.1 Problem Definition

LLM alignment with online human behaviors can be formulated as a Markov Decision Process (MDP) problem, denoted as a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$. We consider human interactions and behaviors as the environment E . At every time step t , the LLM agent observes the current state $s_t \in \mathcal{S}$ from E (the query triggered by the customer) and takes actions $a_t \in \mathcal{A}$ according to a policy $\pi : \mathcal{S} \mapsto p(\mathcal{A})$ that maps the states to

a probability distribution over the actions. The agent’s action is to generate tokens till the end of the content. Then, the agent will receive a reward $r_t = \mathcal{R}(s_t, a_t)$ from the trained reward model, and a new state $s_{t+1} \in \mathcal{S}$ from the environment E . The return of the interactive trajectory $\tau = \{s_1, a_1, \dots, s_T, a_T\}$ is the cumulative γ -discounted rewards, i.e. $R(\tau) = \sum_{t=1}^T \gamma^t r_t$, where T is the horizon of an episode. RL aims to optimize the policy π by maximizing the expected returns from the initial state.

4.2 A Naive Method

Given human behaviors towards <question, LLM response> pairs, one naive method is to directly train feedback simulators to predict user behaviors $b \in \mathcal{B}$ based on the query $s \in \mathcal{S}$ and the LLM response $a \in \mathcal{A}$. Then, the trained simulators can be used as reward models in the RLHF framework to align LLMs with real human behaviors.

In practice, however, collecting sufficient human behaviors related to different answers to the same questions is usually difficult. For example, we may only have one chance to provide the LLM response if users search for some queries that usually never be searched again in a web search engine. Therefore, we propose to build a multi-head pointwise classifier with cross-entropy loss as follows:

$$\text{CE}(\hat{\mathbf{b}}, \mathbf{b}) = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^C b_{ij} \log(\hat{b}_{ij}) \quad (1)$$

After that, the multi-head pointwise classifier is used as the reward model to align LLMs via RLHF algorithm. We thus call it **Reinforcement Learning with Human Behavior through Classifier, RLHBC**.

4.3 Reinforcement Learning with Human Behaviors

Inverse reinforcement learning (IRL) is a popular alternative to align target models with online expert demonstrations, with the advantage of not interacting with experts during training. It enhances sampling and training efficiency, compared to reinforcement learning and imitation learning. Generative Adversarial Imitation Learning (GAIL) (Ho and Ermon, 2016) can further bypass the intermediate step of recovering the unknown reward model in IRL. It directly optimizes the policy, using a GAN discriminator trained by expert demonstrations to provide the action-value function,

$$Q(s, a) = \mathbb{E}_{(s,a) \in \mathcal{M}} [\log(D_\omega(s, a))], \quad (2)$$

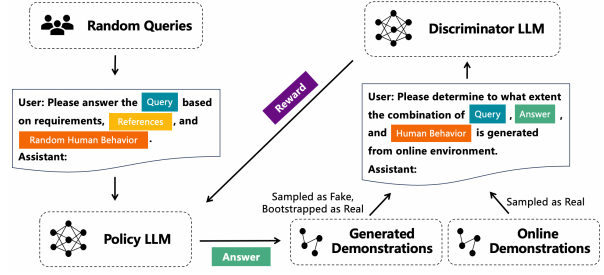


Figure 2: The training process of RLHB.

where the expert and generated demonstrations are denoted as $\mathcal{M}_e$ and $\mathcal{M}_g$. The initial policy actor and discriminator parameters are θ_0 and ω_0 . The discriminator D can be updated by maximizing

$$\mathbb{E}_{(s,a) \in \mathcal{M}_e} [\log(D_\omega(s, a))] + \mathbb{E}_{(s,a) \in \mathcal{M}_g} [1 - \log(D_\omega(s, a))] \quad (3)$$

The discriminator evaluates whether the interaction trajectory comes from expert demonstrations; while the generator is promoted to generate expert-like high-quality content.

In industrial scenarios, online demonstrations with strong positive feedback are considered expert or golden demonstrations. However, strong feedback samples are usually sparse and imbalanced. For example, most search items may receive at most one view or click, but others can attract rich and massive interactions by most users.

Regarding the above problem, we propose a novel alignment method, named **Reinforcement Learning with Human Behaviors (RLHB)**, see Figure 2 for illustration), fully utilizing online demonstrations, inspired by Decision Transformer (Chen et al., 2021). The discriminator here is defined to verify whether the query-answer $\langle s, a \rangle$ pair under the given feedback b comes from real online demonstrations, rather than expert demonstrations.

$$\mathbb{E}_{(s_t, a_t, b_t) \in \mathcal{M}_e} [\log(D(s_t, a_t; b_t))] + \mathbb{E}_{(s_t, a_t, b_t) \in \mathcal{M}_g} [1 - \log(D(s_t, a_t; b_t))] \quad (4)$$

The LLM generator, when receiving a query s , is supposed to give a response a following the expected feedback b , that is, $\pi(a_t | s_t, b_t)$, regardless of positive or negative feedback. When implemented in an online environment, the policy is required to comply with preferred feedback.

We update the generator using the objective function with clipped surrogate following RLHF:

$$\mathcal{L}(\theta) = \mathbb{E}_t \left[\min \left(\ell_t \hat{A}_t, \text{clip}(\ell_t, 1 - \epsilon, 1 + \epsilon) \hat{A}_t \right) \right], \quad (5)$$

where the ratio of the new policy over the old is

$$\ell_t = \frac{\pi_\theta(a_t | s_t, b_t)}{\pi_{\theta_{\text{old}}}(a_t | s_t, b_t)}, \quad (6)$$

and the advantages are calculated following the generalized advantage estimator (GAE):

$$\hat{A}_t = \sum_{l=0}^{\infty} (\gamma \lambda)^l \delta_{t+l}, \quad (7)$$

$$\delta_t = r_t + \gamma V_\phi(s_{t+1}; b_{t+1}) - V_\phi(s_t; b_t), \quad (8)$$

$$\hat{R}_t = \hat{A}_t + V_\phi(s_t; b_t). \quad (9)$$

The 'rewards' r_t is provided by the discriminator $D(s_t, a_t; b_t)$, and $\hat{R}_t$ denotes expected returns. Note that r_t is combined with the KL divergence penalties per token to prevent deviation from the initial model.

$$r_t - \eta \text{KL}(\pi_{\theta_{\text{old}}}(a_t | s_t, b_t), \pi_\theta(a_t | s_t, b_t)) \quad (10)$$

The critic, with parameters ϕ , providing token-level state-values, is updated by

$$\mathcal{L}(\phi) = \mathbb{E}_t [\|V_\phi(s_t; b_t) - \hat{R}_t\|^2]. \quad (11)$$

The parameters of the policy θ , critic ϕ , and discriminator ω models are iteratively updated in sequence during training.

Bootstrap Enhancement Since RLHB introduces simultaneous training of models (i.e. the actor, critic, and discriminator), the efficiency and balance of sampling are crucial to each. The outputs of the critic and discriminator are both scalar values, which are relatively easy to learn; in contrast, the convergence of the actor is more difficult owing to larger and more uncertain action space. To prevent the discriminator from overfitting and make it more robust, we bootstrap a certain proportion κ of highly-rewarded demonstrations as 'fake' online demonstrations for discriminator updating.

Behavior in Natural-Language Form Furthermore, we change the form of behavioral data modeling. In RLHF-type methods, data processes in pairwise or numerical forms are necessary before reward modeling and RL alignment, breaking the training process into parts and bringing instability in signal models. However, LLM has proven its ability to understand and express natural language like humans, including text containing data and statistical results. Therefore, under the settings of RLHB, we instead utilize the intrinsic properties of LLM to describe human behaviors in the form of natural language and put it into instructions.

5 Experiments

In this section, we conduct several experiments to assess the effectiveness of our methods, which aims to answer the following two questions:

Q1 Whether the generative model can be directly aligned on real online human behaviors?

Q2 Whether the predefined preferences alignment can be further improved by online alignment?

5.1 Data

Human Behavior Data. Online human behavior data is collected from Baidu Search. We uniformly sample real demonstrations from online environments based on indicators discussed in Section 3. We finally obtained around 100k <query, answer, feedback> triplets for experiments.

Human Preference Data. We also collect manual preference data for reward modeling, following InstructGPT (Ouyang et al., 2022). We sample random queries and generate multiple answers for each using LLMs for internal usage. Then, every two different answers to the same query are annotated with preference labels by experienced annotators¹.

5.2 Models

Baseline (SFT). We utilize an internal LLM (13B) as the backbone and set its default version as the baseline, fine-tuned well with millions of platform-owned data.

Reward Model (RM). RM is trained on human preference data with pairwise ranking loss in Eq. 12. ψ denotes the RM parameters, and a_w is the preferred answer for each pair.

$$\mathcal{L}(\psi) = -\log \sigma(r_\psi(s, a_w) - r_\psi(s, a_l)) \quad (12)$$

Classifier Model (CM). Following setups in Section 4.2 and 5.1, the classifier in RLHBC fits human behavior data in numerical form.

Discriminator Model (DM). We also warm up a discriminator for ablation experiments. It is trained using online human behavior samples as real, and random replacements of their feedback as fake.

¹The annotation consistency confidence level is over 80%

5.3 Evaluation Metrics

Three types of metrics are used for evaluation. First, as GPT4 (OpenAI, 2023) is widely used as a reference to evaluate the quality of generated text (Wang et al., 2023), we follow these studies to prompt GPT4² to rank each pair of responses to the same query from different models and then summarize them into Win-Tie-Loss (WTL) results. Moreover, we adopt two human evaluation systems in Baidu Search to provide fine-grained assessments, i.e. Quality Score and Satisfaction Score. We recruit annotators, well-educated and experienced in the labeling criteria, to label these scores for each pair of queries and LLM responses. Additionally, following other studies on RL-based alignment methods (Zheng et al., 2023), we utilize RL indicators to evaluate the stability of model training and LLM metrics to assess the LLM capabilities.

Quality Score. It focuses on content quality with four levels. **Bad** indicates text with low quality, mistakes, irrelevant or false contents; **Medium** means redundant texts or those that do not solve the major need; **Good** denotes the text can meet the need with the lack of content richness, relevance, authoritativeness, or necessary multi-modal information; **Excellent** represents the generated text fully and accurately meets needs and adds in-depth and extended content. Based on Quality Scores, pairwise Win-Tie-Loss can also be provided.

Satisfaction Score. It is a three-level score to measure user satisfaction with given responses to their search queries. **Unsatisfied** mainly indicates content quality problems, such as irrelevance, mistakes, duplications, hallucinations, partial satisfaction of the user’s need, and some other defects; **Partially Satisfied** means the answers meet main requirements but in an unsatisfied display form, such as under-listing, over-listing, excessive length, inappropriate integration of multi-modal contents, etc.; **Satisfied** demonstrates that satisfied or even in-depth content is integrated into a proper format and length, without any wrong information.

5.4 Other Setups

For Q1, we trained RLHF, RLHBC, and RLHB from the baseline to see which produces more alignment gains for the unaligned SFT. The critic models for RLHF and RLHBC are started from RM and CM; while the critic and discriminator models for

RLHB are initialized from the baseline. For Q2, we further train RLHBC and RLHB based on RLHF to see if additional improvements can be made, denoted as RLHF + RLHBC and RLHF + RLHB.

6 Experimental Results

6.1 GPT4 Evaluation

We obtain generated responses to random queries and combine them into pairs. Then, we utilize evaluation prompts to ask GPT4 to select a better response. Note that we combine each pair in forward and reverse orders and ask GPT4 to rank them twice. The results are shown in Figure 3(a).

In the quantity of Win-Tie-Loss, the RLHBC and RLHB perform better than their baseline SFT; while the RLHF + RLHBC and RLHF + RLHB are worse than their baseline RLHF. However, based on the Sign Test, the performance of these four groups is statistically equal. Moreover, it is noteworthy that, compared to SFT, RLHF shows significant improvement, with a win rate of 71%. In brief, from the perspective of GPT4, regardless of the baselines, the improvement brought by online interactive alignment is limited and may even weaken the performance in some cases.

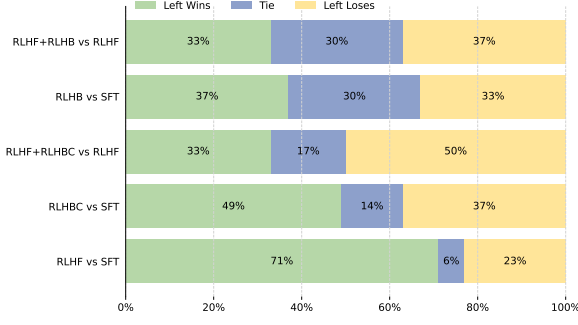
However, studies found that the limitations and knowledge barriers of GPT4 may impair its reliability of results (Wang et al., 2023). Thus, we further leverage manual evaluation systems to provide a comprehensive assessment.

6.2 Human Evaluation

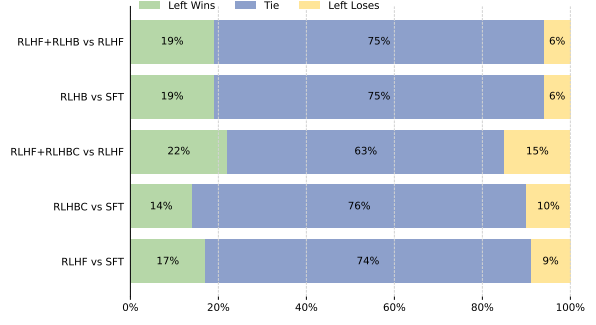
Figure 3(c) demonstrates human evaluation results by Quality Scores, and we mainly focus on Excellent and Good improvements. First, all candidates show improvements over the baseline SFT, ranging from 6% to 18%. The performance of RLHB and RLHF is basically the same, indicating question Q1 is feasible. Yet, RLHBC is not better than RLHF, which may result from the Classifier being sensitive to the training data distribution. Furthermore, RLHF + RLHBC and RLHF + RLHB achieve improvements of 3% to 7% over their baseline RLHF. It indicates that question Q2 is also workable, enabling further gains from online interactive alignment based on predefined preference alignment. On the whole, RLHF + RLHB is the only one that achieves 60% high-quality ratio, satisfying users’ requirements without hallucinations or defects.

Besides, we report Win-Tie-Loss results according to Quality Scores in Figure 3(b). Different from

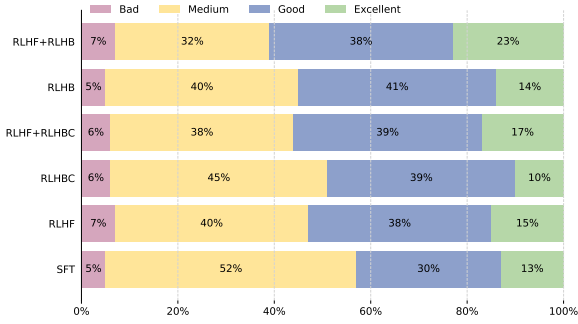
²GPT4-Turbo is used in our evaluation.



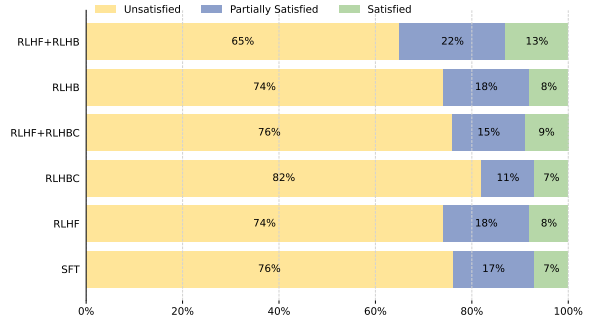
(a) GPT-4 evaluation over quality WTL.



(b) Human evaluation over quality WTL.



(c) Human evaluation over quality score.



(d) Human evaluation over satisfaction score.

Figure 3: GPT4 and Human Evaluations. Even though the results differ to varying degrees, most confirm the feasibility of the proposed questions Q1 and Q2, especially from the perspective of RLHB and RLHF + RLHB.

GPT4’s results, the performance of RLHF and SFT is statistically equal here. Additionally, with a rate of Win-Loss 19%:6%, RLHF + RLHB and RLHB present significant improvements from RLHF and SFT, respectively. It confirms the feasibility of the two questions again. On the contrary, the enhancement of RLHF + RLHBC and RLHBC over RLHF and SFT is just near statistical significance. Another problem occurs to RLHF + RLHBC. When the win rate increases to 22%, the loss rate also rises to 15%, both the highest among all comparisons. Therefore, RLHB-based models are more promising in industrial applications.

Figure 3(d) displays the outcomes of Satisfaction Scores, and we mainly focus on the percentages of Satisfied and Partially Satisfied. Firstly, RLHB and RLHF perform similarly; while RLHF + RLHB further improves RLHF by 9%. These once again confirm the two questions are feasible. However, both RLHF + RLHBC and RLHBC are worse than their baselines RLHF and SFT, respectively. It reflects the stability issues of Classifier-based alignment models. Interestingly, on Satisfaction dimensions, the effects of RLHF and SFT are still the same, which is quite different from GPT4’s results.

6.3 Model Training Metrics

In this subsection, we study the influences of hyper-parameter setups on model performances. The training metrics for RLHF, RLHBC, and RLHF + RLHBC are presented in Figure 4. It is worth noting that the win rates by RM and mean rewards for RLHF present obvious upward trends, approaching 0.95 and 3.0 respectively. In contrast, for RLHBC and RLHF + RLHBC, the win rates fluctuate around 0.4, and the mean rewards reach 2.25 only. Compared with other methods, RLHF achieves better convergence under the incentive of RM.

The training metrics for RLHB and RLHF + RLHB are shown in Figure 5. We can see that the discriminator loss tends to converge stably. However, the rewards of discriminators show different trends. Starting around 0.5, those reward scores decline sharply while the discriminative abilities increase. As the policy is optimized, the rewards decline to around 0.4 in contrast, showing the discriminator is confused to some extent. Yet, the instability of discriminators affects the critic model fitting, leading to fluctuations in expected returns.

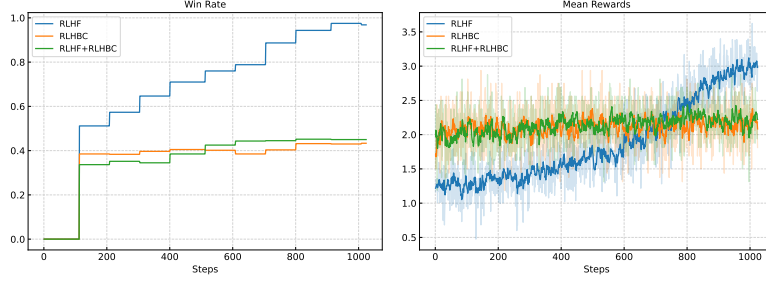


Figure 4: Win Rates and Mean Rewards for RLHF, RLHBC, and RLHF + RLHBC. Compared with CM of RLHBC models, RM of RLHF is much easier to guide the model to achieve preference learning and model convergence.

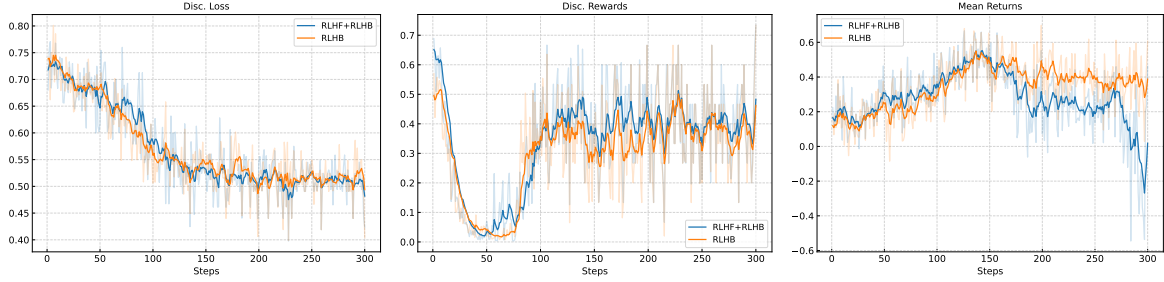


Figure 5: Discriminator Loss, Discriminator Rewards, and Mean Returns for RLHB and RLHF + RLHB. Different from before, the rewards in RLHB will not continue to grow, but will eventually converge to a confusion state, close to 0.5, instead. That means the policy generation ability can already confuse the real with the fake, though this instability may lower the expected returns.

6.4 Ablation Study

In Figure 6, we conduct several ablation experiments. As for the discriminator, we set κ to different values ranging from 0% to 50%, to study the model sensitivity to the proportion of fake-real data, as discussed in Section 4.3. In addition, we also experiment on the trained discriminator in Section 5.2 and freeze its parameters during RLHB training, to see if the discriminator can be trained and used in the way of the reward model. As for the actor (i.e., the LLM policy agent), we compare different numbers of rollouts and batch sizes, set to 4 and 16 by default respectively, to see the impact of the sampling scale.

For the discriminator, as shown in Figure 6(a) and 6(b), when the bootstrap proportion κ reaches 50%, the discriminator collapses sharply, with the rewards approach almost 1.0. Moreover, the frozen discriminator tends to predict the generated samples at a stable score of around 0.6. It cannot be observed that the impact of improved policy generation capabilities on the discriminator. Thus, without the adversarial mechanism, the discriminator cannot work like RM. For the actor, see Figure 6(c), with rollouts increasing from 1 to 4, and 6, the

stability can be significantly improved, especially from mean returns. By changing the batch size 16 to 8, the variance in training effect is not obvious, except for a slightly larger volatility.

7 Conclusions and Discussion

In this work, we explore aligning LLM with the behavioral preferences from online users. We propose two alignment methods, RLHBC and RLHB, and consider online behavior in both natural language form and scalar form. Numerous experiments show that: On the one hand, the LLM alignment based on online human behavior can approximate the alignment based on offline annotated preferences; On the other hand, it can also be concluded that online behavior alignment can further enhance the LLM aligned with offline preferences.

RLHB exhibits numerous advantages compared to RLHF and RLHBC. Whether it is RLHF or RLHBC, the reward model or classifier model must be trained with preprocessed scalar-form human behavior before RL alignment, thus preventing the model from actively and continuously learning online. On the contrary, RLHB can realign the policy model immediately once human feedbacks are

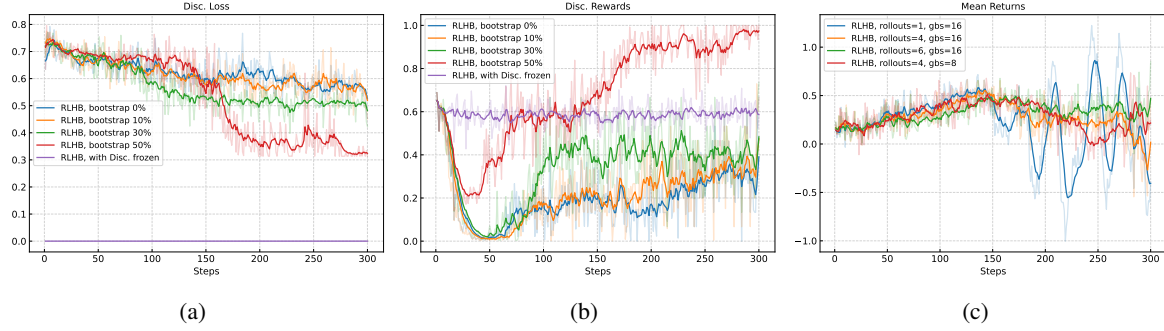


Figure 6: The Ablation Experiments for RLHB.

updated, owing to its multi-model joint training mechanism and natural language-style signal data modeling method. Furthermore, online interaction process is often multi-turn and context-dependent. Either pairwise-based reward models in RLHF or pointwise-based classifier models in RLHBC can only model one-round interaction, hindering scalable and sustainable training for online alignment. However, the IRL mechanism behind RLHB and the natural language-based behavioral modeling provide room for solving this problem, which can also inspire future research.

References

- Anthropic and the Collective Intelligence Project. 2023. [Collective constitutional ai: Aligning a language model with public input](#).
- Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. 2023. A general theoretical paradigm to understand learning from human preferences. *arXiv preprint arXiv:2310.12036*.
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.
- Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*.
- Jonathan D Chang, Kianté Brantley, Rajkumar Ramamurthy, Dipendra Misra, and Wen Sun. 2023. Learning to generate better than your llm. *arXiv preprint arXiv:2306.11816*.
- Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. 2021. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097.
- Hanze Dong, Wei Xiong, Deepanshu Goyal, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. 2023a. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*.
- Yi Dong, Zhilin Wang, Makesh Narsimhan Sreedhar, Xianchao Wu, and Oleksii Kuchaiev. 2023b. Steerlm: Attribute conditioned sft as an (user-steerable) alternative to rlhf. *arXiv preprint arXiv:2310.05344*.
- Yang Gao, Dana Alon, and Donald Metzler. 2024. Impact of preference noise on the alignment performance of generative language models. *arXiv preprint arXiv:2404.09824*.
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.
- Jonathan Ho and Stefano Ermon. 2016. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29.
- Jian Hu, Li Tao, June Yang, and Chandler Zhou. 2023. Aligning language models with offline reinforcement learning from human feedback. *arXiv preprint arXiv:2308.12050*.
- Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE international conference on data mining*, pages 263–272. Ieee.
- Jui-Ting Huang, Ashish Sharma, Shuying Sun, Li Xia, David Zhang, Philip Pronin, Janani Padmanabhan, Giuseppe Ottaviano, and Linjun Yang. 2020. Embedding-based retrieval in facebook search. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2553–2561.

- Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. 2023. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*.
- Di Jin, Shikib Mehri, Devamanyu Hazarika, Aishwarya Padmakumar, Sungjin Lee, Yang Liu, and Mahdi Namazifar. 2023. Data-efficient alignment of large language models with human feedback through natural language. *arXiv preprint arXiv:2311.14543*.
- Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2017. Accurately interpreting clickthrough data as implicit feedback. In *Acm Sigir Forum*, volume 51, pages 4–11. Acm New York, NY, USA.
- Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. 2023. [Evaluating open-domain question answering in the era of large language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606, Toronto, Canada. Association for Computational Linguistics.
- Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbone, and Abhinav Rastogi. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.
- Qianxi Li, Yingyue Cao, Jikun Kang, Tianpei Yang, Xi Chen, Jun Jin, and Matthew E Taylor. 2023a. Laffi: Leveraging hybrid natural language feedback for fine-tuning language models. *arXiv preprint arXiv:2401.00907*.
- Yuhui Li, Fangyun Wei, Jinjing Zhao, Chao Zhang, and Hongyang Zhang. 2023b. Rain: Your language models can align themselves without finetuning. *arXiv preprint arXiv:2309.07124*.
- Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let’s verify step by step. *arXiv preprint arXiv:2305.20050*.
- Nathan N Liu, Evan W Xiang, Min Zhao, and Qiang Yang. 2010. Unifying explicit and implicit feedback for collaborative filtering. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1445–1448.
- Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583*.
- Bhaskar Mitra, Nick Craswell, et al. 2018. An introduction to neural information retrieval. *Foundations and Trends® in Information Retrieval*, 13(1):1–126.
- Andrew Y Ng, Stuart Russell, et al. 2000. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, page 2.
- OpenAI. 2023. [Gpt-4 technical report](#).
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.
- Yujia Qin, Zihan Cai, Dian Jin, Lan Yan, Shihao Liang, Kunlun Zhu, Yankai Lin, Xu Han, Ning Ding, Huadong Wang, Ruobing Xie, Fanchao Qi, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2023. WebCPM: Interactive web search for Chinese long-form question answering. In *ACL*, pages 8968–8988.
- 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*.
- Alexandre Rame, Guillaume Couairon, Mustafa Shukor, Corentin Dancette, Jean-Baptiste Gaya, Laure Soulier, and Matthieu Cord. 2023. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. *arXiv preprint arXiv:2306.04488*.
- Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *ArXiv*, abs/2302.04761.
- Taiwei Shi, Kai Chen, and Jieyu Zhao. 2023. Safer-instruct: Aligning language models with automated preference data. *arXiv preprint arXiv:2311.08685*.
- 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.
- Zhiqing Sun, Yikang Shen, Hongxin Zhang, Qinzhong Zhou, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2023. Salmon: Self-alignment with principle-following reward models. *arXiv preprint arXiv:2310.05910*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrutika Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

- Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Cl  mentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.
- 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*.
- Tianlu Wang, Ping Yu, Xiaoqing Ellen Tan, Sean O’Brien, Ramakanth Pasunuru, Jane Dwivedi-Yu, Olga Golovneva, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023. Shepherd: A critic for language model generation. *arXiv preprint arXiv:2308.04592*.
- Chuhan Wu, Fangzhao Wu, Tao Qi, Qi Liu, Xuan Tian, Jie Li, Wei He, Yongfeng Huang, and Xing Xie. 2022. Feedrec: News feed recommendation with various user feedbacks. In *Proceedings of the ACM Web Conference 2022*, pages 2088–2097.
- Zequ 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*.
- Ruobing Xie, Cheng Ling, Yalong Wang, Rui Wang, Feng Xia, and Leyu Lin. 2021. Deep feedback network for recommendation. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*, pages 2519–2525.
- Wenda Xu, Daniel Deutsch, Mara Finkelstein, Juraj Juraska, Biao Zhang, Zhongtao Liu, William Yang Wang, Lei Li, and Markus Freitag. 2023. Pinpoint, not criticize: Refining large language models via fine-grained actionable feedback. *arXiv preprint arXiv:2311.09336*.
- Kevin Yang, Dan Klein, Asli Celikyilmaz, Nanyun Peng, and Yuandong Tian. 2023. Rlcd: Reinforcement learning from contrast distillation for language model alignment. *arXiv preprint arXiv:2307.12950*.
- Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Chuanqi Tan, and Chang Zhou. 2023a. Scaling relationship on learning mathematical reasoning with large language models. *arXiv preprint arXiv:2308.01825*.
- Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2023b. Rrhf: Rank responses to align language models with human feedback without tears. *arXiv preprint arXiv:2304.05302*.
- Dun Zeng, Yong Dai, Pengyu Cheng, Tianhao Hu, Wanshun Chen, Nan Du, and Zenglin Xu. 2023. On diverse preferences for large language model alignment. *arXiv preprint arXiv:2312.07401*.
- Qian Zhao, F Maxwell Harper, Gediminas Adomavicius, and Joseph A Konstan. 2018. Explicit or implicit feedback? engagement or satisfaction? a field experiment on machine-learning-based recommender systems. In *Proceedings of the 33rd Annual ACM symposium on applied computing*, pages 1331–1340.
- Rui Zheng, Shihan Dou, Songyang Gao, Yuan Hua, Wei Shen, Binghai Wang, Yan Liu, Senjie Jin, Qin Liu, Yuhao Zhou, et al. 2023. Secrets of rlhf in large language models part i: Ppo. *arXiv preprint arXiv:2307.04964*.
- Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji-Rong Wen. 2023. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*.