

---

# Trustworthy Alignment of Retrieval-Augmented Large Language Models via Reinforcement Learning

---

Zongmeng Zhang<sup>1</sup> Yufeng Shi<sup>1</sup> Jinhua Zhu<sup>1</sup> Wengang Zhou<sup>1,2</sup> Xiang Qi<sup>3</sup> Peng Zhang<sup>3</sup> Houqiang Li<sup>1,2</sup>

## Abstract

Trustworthiness is an essential prerequisite for the real-world application of large language models. In this paper, we focus on the trustworthiness of language models with respect to retrieval augmentation. Despite being supported with external evidence, retrieval-augmented generation still suffers from hallucinations, one primary cause of which is the conflict between contextual and parametric knowledge. We deem that retrieval-augmented language models have the inherent capabilities of supplying response according to both contextual and parametric knowledge. Inspired by aligning language models with human preference, we take the first step towards aligning retrieval-augmented language models to a status where it responds relying merely on the external evidence and disregards the interference of parametric knowledge. Specifically, we propose a reinforcement learning based algorithm TRUSTWORTHY-ALIGNMENT, theoretically and experimentally demonstrating large language models' capability of reaching a trustworthy status without explicit supervision on how to respond. Our work highlights the potential of large language models on exploring its intrinsic abilities by its own and expands the application scenarios of alignment from fulfilling human preference to creating trustworthy agents. Our code is available at <https://github.com/zmzhang2000/trustworthy-alignment>.

## 1. Introduction

Recent advancement in Large Language Models (LLMs) has significantly boosted the application of Artificial Intelligence (AI) systems across a wide range of fields, such as dialogue assistants (OpenAI, 2023) and autonomous agents (Wang et al., 2023a). Researchers contend that one of the most powerful LLMs, *i.e.*, GPT-4, could reasonably be viewed as a spark of Artificial General Intelligence (AGI) considering its capabilities in language, vision, mathematics, coding, *etc* (Bubeck et al., 2023).

Despite these remarkable capabilities, the lack of trustworthiness is still a challenging obstacle to the application of LLMs. As probabilistic models, LLMs occasionally generate output which is plausible but deviating from source input (Adlakha et al., 2023) or factual knowledge (Min et al., 2023; Muhlgay et al., 2023; Li et al., 2023a). This phenomenon is referred to as *hallucination* (Maynez et al., 2020; Ji et al., 2023b), which poses tangible risks to real-world application scenarios with low tolerance of error such as medical diagnosis (Nori et al., 2023) and legal consultation (Cui et al., 2023; Nguyen, 2023).

A common and straightforward approach to mitigate hallucinations in LLMs is to leverage non-parametric knowledge source through *Retrieval-Augmented Generation* (RAG) (Lewis et al., 2020; Shuster et al., 2021). Specifically, RAG retrieves supporting evidence relevant to the source input from an external knowledge base through a retriever, and generates the output conditioning on both the source input and retrieved evidence. The retrieve-and-generate paradigm proves to significantly enhance the factuality and reliability of LLMs (Lewis et al., 2020; Borgeaud et al., 2021; Lazaridou et al., 2022).

However, incorporating retrieval into language generation encounters a new problem of *knowledge conflicts* (Longpre et al., 2021) that still impedes the progress towards trustworthy language models. LLMs commonly learn from massive unsupervised data during pre-training and store the learned knowledge in their parameters. Conflict emerges as contextual knowledge in the retrieved evidence contrasting with the inner knowledge stored in LLMs' parameters, which triggers "hallucinations" if an LLM holds a strong

---

<sup>1</sup>University of Science and Technology of China <sup>2</sup>Institute of Artificial Intelligence, Hefei Comprehensive Nation Science Center <sup>3</sup>Ant Group. Correspondence to: Wengang Zhou <zhwg@ustc.edu.cn>, Houqiang Li <lihq@ustc.edu.cn>.

belief in its parametric knowledge and produces response inconsistent with the provided context. This phenomenon is unexpected as the performance of RAG will be restricted by hallucinations if the retrieved evidence is not accurate and helpful enough (Karpukhin et al., 2020; Shi et al., 2023; Ren et al., 2023; Mallen et al., 2023). We focus our research on the retrieval-augmented generation setting where an LLM is deemed trustworthy if it answers questions based on contextual rather than parametric knowledge.

A wealth of studies (Longpre et al., 2021; Chen et al., 2022; Neeman et al., 2023; Zhou et al., 2023) has been dedicated to this problem owing to its obstructive impact on trustworthy LLMs. These efforts have discovered LLMs’ dilemma between the parameters and retrieved evidence. In other words, LLMs may have capabilities of responding according to both contextual and parametric knowledge, and struggles to determine which to rely on. The inherent capabilities conforms to the foundation of a recently emerged technology in the field of AI: *alignment* (Ji et al., 2023a), which aims at encouraging a instruction-following language model (Wei et al., 2022) to behave in line with human intentions and values (Leike et al., 2018) by reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022). It is worth noting that alignment only “unlocks” capabilities LLMs already had, but does not teach new capabilities to the pre-trained and instruction-tuned models (Lowe & Leike, 2022).

Inspired by the success of aligning LLMs with human intentions and values, we take the first step towards aligning retrieval-augmented LLM to a trustworthy status where it supplies responses in accordance with merely the retrieved evidence and ignores the parametric knowledge. We hypothesize that LLMs have tendencies of generating response in line with both contextual and parametric knowledge. Based on this, we address the problem of knowledge conflicts by aligning retrieval-augmented LLMs to trustworthy status via reinforcement learning (**TRUSTWORTHY-ALIGNMENT**). In particular, we elaborate our work by answering the following research questions:

- ( $RQ_1$ ) *Does the hypothesis hold that retrieval-augmented LLMs have inherent capabilities of responding according to both contextual and parametric knowledge?* (Section 2)
- ( $RQ_2$ ) *How can we design an algorithm that enhances the trustworthiness of retrieval-augmented LLMs by aligning them to a trustworthy status via reinforcement learning?* (Section 3)
- ( $RQ_3$ ) *How does the proposed algorithm perform and what are the side effects of aligning retrieval-augmented LLMs to trustworthy status?* (Section 4)

We demonstrate that when *merely providing what to achieve*, retrieval-augmented LLMs are capable of aligning itself to

a status where it furnishes responses only relying on the contextual knowledge *without explicit supervision on how to respond*. Our work highlights the potential of LLMs on exploring its intrinsic capabilities by its own and expands the application scenarios of alignment from fulfilling human intention and values to creating trustworthy agents.

## 2. Preliminary Study

In this section, we first provide the background of **TRUSTWORTHY-ALIGNMENT** by formulating the task of retrieval-augmented generation, and discuss the problem of knowledge conflicts. Subsequently, we introduce answer probability, the proxy metric for model tendency analysis. Finally, we present the model tendency under different human-curated prompts, answering the prerequisite question of our work: ( $RQ_1$ ) *does the hypothesis hold that retrieval-augmented LLMs have inherent capabilities of responding according to both contextual and parametric knowledge?*

### 2.1. Retrieval-Augmented Generation

A language model is a parametric probabilistic model  $\pi_\theta$  conditioned on the input text. This paper focuses on the question answering (QA) task where the input text is a question  $q$ . The generation process is then formulated as sampling a response  $y$  from the language model:

$$y \sim \pi_\theta(y|q) = \prod_{t=0}^{T-1} \pi_\theta(y_t|q, y_{<t}), \quad (1)$$

where both  $x$  and  $y$  are sequences of tokens,  $t$  is the token index, and  $T$  is the token number of the response.

Under the retrieval-augmented setting, we have an additional context  $c$  as a supporting evidence for answering the question, which is typically retrieved from a large knowledge base, e.g., Wikipedia. Following Brown et al. (2020), we formulate the retrieval-augmented generation task by reformatting the question  $q$  and context  $c$  as a natural language instruction using a prompting function  $f(q, c)$ :

$$y \sim \pi_\theta(y|q, c) = \prod_{t=0}^{T-1} \pi_\theta(y_t|f(q, c), y_{<t}). \quad (2)$$

### 2.2. Knowledge Conflicts

Ideally, a retrieval-augmented LLM responds faithfully in line with the knowledge from  $c$ . However, the reality falls short of expectations since we have no idea whether the contextual knowledge in  $c$  or the parametric knowledge in  $\theta$  contributes more to LLM’s decision. The *knowledge conflict* (Longpre et al., 2021) emerges when contextual knowledge diverges from the parametric knowledge.

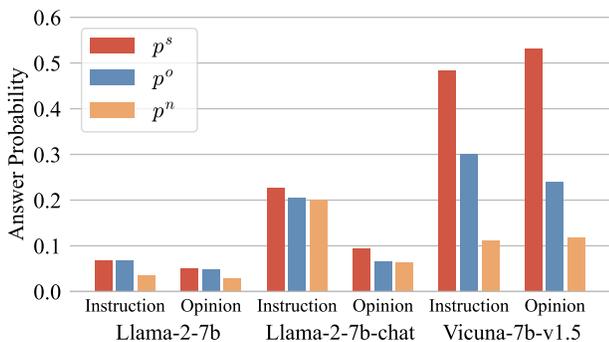


Figure 1. Average answer probabilities of the substituted answer  $p^s$ , original answer  $p^o$  and none response  $p^n$  for Llama-2-7b, Llama-2-7b-chat (Touvron et al., 2023) and Vicuna-7b-v1.5 (Zheng et al., 2023) with instruction-based and opinion-based prompting strategies. Generating tokens other than option codes is also feasible so that the sum of three probabilities is less than 1. Details of prompts are listed in Appendix E.

We simulate knowledge conflicts by constructing counterfactual data following the entity substitution framework in Longpre et al. (2021). Specifically, we have an original triplet  $(q, c^o, y^o)$ , where  $y^o$  is the original right answer of question  $q$  and is included in the context  $c^o$  so that it can be inferred from  $q$  and  $c^o$ . For each sample, we detect the corresponding entity of  $y^o$  and replace it with a another entity  $y^s$  of the same type. Meanwhile,  $c^o$  is transformed to a new context  $c^s$  by replacing all of the answer entity mentions. We expect the language model to supply a response including the substituted answer  $y^s$  given the counterfactual context  $c^s$ .

### 2.3. Tendency Analysis

To validate the prerequisite of our work that retrieval-augmented LLMs have tendencies of delivering response according to both the contextual and parametric knowledge, we conduct tendency analysis on multiple prevalent pre-trained LLMs. For a specific question  $q$  and the counterfactual context  $c^s$ , we construct three candidate responses: the substituted answer  $y^s$ , the original answer  $y^o$  and a none response  $y^n$  instantiated as “None of the above”. Then we format the question answering problem to a multi-choice task where LLM selects the best response from  $y^s$ ,  $y^o$  and  $y^n$  given  $q$  and  $c^s$ . We propose *answer probability* as a proxy for measuring whether the model relies more on contextual or parametric knowledge, computed by

$$\pi_\theta([\text{OPTION}]|f_{\text{multi-choice}}(q, c^s, y^s, y^o, y^n)), \quad (3)$$

where [OPTION] is the token of option code, *i.e.*, [A], [B] or [C] in out setting. We denote the probabilities of  $y^s$ ,  $y^o$  and  $y^n$  corresponding to the options as  $p^s$ ,  $p^o$ ,  $p^n$ , respectively.

We visualize the tendency of three pre-trained LLMs with two types of prompting strategies on the test set of Natural Questions (Fisch et al., 2019) in Figure 1. With both the instruction-based and opinion-based prompts, Llama-2-7b and Llama-2-7b-chat give almost the same probabilities for the substituted and original answer. Vicuna-7b-v1.5 shows maximum sum of the three probabilities, indicating strongest understanding of the multi-choice task. Even so, Vicuna-7b-v1.5 struggles to decisively choose the substituted answer and still has an likelihood around 0.25 to select the original answer. Hence we draw the conclusion that *the hypothesis of LLMs having capabilities of responding according to both the contextual and parametric knowledge holds true.*

## 3. Trustworthy Alignment

The inherent capabilities of responding according to the contextual and parametric knowledge afford the opportunity to control the LLMs’ intention in the scenario of knowledge conflict. In this section, we address  $(RQ_2)$  by *formulating the LLM as a Markov Decision Process (MDP), introducing TRUSTWORTHY-ALIGNMENT and providing theoretical analysis for the proposed algorithm.*

### 3.1. LLM as an MDP

Typically, an MDP  $\mathcal{M} \triangleq \langle \mathcal{S}, \mathcal{A}, r, \mathbb{P}, \rho_0, \gamma \rangle$  consists of the state space  $\mathcal{S}$ , the action space  $\mathcal{A}$ , a reward function  $r$ , the transition probability  $\mathbb{P}$ , the initial state distribution  $\rho_0$  and a discount factor  $\gamma$ . The generation process of an LLM is naturally an MDP where each state  $s \in \mathcal{S}$  is the context comprising prompt tokens  $\tilde{a}_{0:m}$  and previously generated tokens  $a_{0:t-1}$ . The initial state  $\tilde{s}$  is the sequence of prompt tokens  $\tilde{a}_{0:m}$  sampled from the distribution  $\rho_0$ .<sup>1</sup> Action space  $\mathcal{A}$  is a finite vocabulary  $\mathcal{V} = \{1, \dots, |\mathcal{V}|\}$ . Given the current state  $s_t = a_{0:t-1}$ , the LLM selects an action  $a_t$  from  $\mathcal{A}$ , identically samples a token from the vocabulary  $\mathcal{V}$  with the distribution  $\pi_\theta$  in an auto-regressive manner formulated as  $a_t \sim \pi_\theta(a_t|s_t) = \pi_\theta(a_t|a_{0:t-1})$ . The transition in LLM is deterministic as the next state  $s_{t+1}$  is definitely transformed into  $a_{0:t}$  after the action  $a_t$  is selected. Afterward, a deterministic reward  $r_t$  is given by the reward function:  $r_t = r(s_t, a_t)$ . A state is regarded as terminated when a special end-of-text token is selected or the response reaches the max length. We define the state action value function  $Q^{\pi_\theta}(s, a) \triangleq \mathbb{E}_{\mathbb{P}, \pi_\theta} \left[ \sum_{t=0}^{T-1} \gamma^t r^t | s_0 = s, a_0 = a \right]$ , which represents the expected return following policy  $\pi_\theta$  after taking action  $a$  at the state  $s$ . The state value function is defined as  $V^{\pi_\theta} = \mathbb{E}_{a \sim \pi_\theta} [Q^{\pi_\theta}(s, a)]$ , *i.e.*, the expected

<sup>1</sup>As all of the subsequent generation steps condition on the prompt tokens, we omit  $\tilde{a}_{0:m}$  and denote state  $s_t$  only with the generated tokens as  $a_{0:t-1}$  for clarity.

return following  $\pi_\theta$  from state  $s$ . The primary objective of optimizing the LLM via reinforcement learning is to maximize the expectation of discounted cumulative reward:

$$J(\pi_\theta) = \mathbb{E}_{\tilde{s} \sim \rho_0} \mathbb{E}_{a_{0:T-1} \sim \pi_\theta} \left[ \sum_{t=0}^{T-1} \gamma^t r_t \right]. \quad (4)$$

### 3.2. Objective of Trustworthy-Alignment

We aim at aligning the language model to the trustworthy status where it supplies responses in accordance with merely the retrieved evidence and ignores the parametric knowledge. To achieve this goal, we propose a simple yet effective reinforcement learning (RL) algorithm which rewards the model’s reliance on contextual knowledge and penalizes its reliance on parametric knowledge. Specifically, given the question  $q$  and substituted context  $c^s$ , we reward  $\pi_\theta$  as long as the counterfactual answer  $y^s$  appears in the terminated response and penalize it if the original answer  $y^o$  appears or neither answers are included, formulated as

$$r_t^{\text{trust}} = \begin{cases} +\alpha_0 & \text{if } s_{t+1} \text{ ends } \wedge y^s \subseteq s_{t+1} \wedge y^o \not\subseteq s_{t+1} \\ -\alpha_1 & \text{if } s_{t+1} \text{ ends } \wedge y^s \not\subseteq s_{t+1} \wedge y^o \not\subseteq s_{t+1} \\ -\alpha_0 & \text{if } s_{t+1} \text{ ends } \wedge y^o \subseteq s_{t+1} \\ 0 & \text{otherwise} \end{cases}, \quad (5)$$

where  $\alpha_0$  and  $\alpha_1$  control absolute values of reward and penalty. Note that only the last generated token is rewarded so that this reward is on the sentence level.

Indispensably, a Kullback–Leibler (KL) divergence term is added to the reward following preceding works (Jaques et al., 2019; Stiennon et al., 2020) to penalize the policy’s deviation from the original supervised fine-tuned model  $\pi^{\text{SFT}}$ , formulated as

$$r_t^{\text{KL}} = -\alpha_2 \log \left[ \frac{\pi_\theta(a_t|s_t)}{\pi^{\text{SFT}}(a_t|s_t)} \right], \quad (6)$$

where  $\alpha_2$  is a coefficient controlling the training trade-off between a model performing specific task and a model more like the original one.

Even with the regularization of KL divergence, LLMs are prone to falling into local optima in the training procedure of RL. For instance, pre-trained LLM may continually output the same repeated pattern when a short repetitive string like “\n\n\n” shows up in the prompt, which will be reinforced and overwhelm the cumulative  $r_t^{\text{trust}}$  in the subsequent training procedure by virtue of its low accumulative KL penalty. To avoid this, we propose an additional collapse penalty to persuade the LLM from collapsed mode:

$$r_t^{\text{col}} = g(s_{t+1}) = \begin{cases} -\alpha_3 & \text{if } s_{t+1} \text{ ends and collapses} \\ 0 & \text{otherwise} \end{cases}, \quad (7)$$

where  $\alpha_3$  controls the absolute value of  $r_t^{\text{col}}$  and the collapse detection algorithm is described in Appendix C.

Overall, the final reward function is formulated as

$$r(s_t, a_t) = r_t^{\text{trust}} + r_t^{\text{col}} + r_t^{\text{KL}}. \quad (8)$$

We train the policy with PPO algorithm (Schulman et al., 2017). We propose an annealing strategy to trade-off between exploration and exploitation in PPO training of LLMs, which is detailed in Appendix B due to space constraints. The algorithm pipeline of TRUSTWORTHY-ALIGNMENT is described in Appendix C.

### 3.3. Theoretical Analysis

In this subsection, we theoretically analyze the proposed TRUSTWORTHY-ALIGNMENT, ensuring the policy’s convergence to the trustworthy status. Specifically, we demonstrate that the optimal policy for the total reward  $r(s_t, a_t)$  is consistent with the optimal policy for the task-specific reward  $r_t^{\text{trust}}$ , while  $r_t^{\text{KL}}$  and  $r_t^{\text{col}}$  do not affect the realization of the alignment. Actually, the purpose of designing these regular terms is to accelerate the model’s learning procedure in practice (Ng et al., 1999).

Formally, we define an original MDP as  $\mathcal{M} \triangleq \langle \mathcal{S}, \mathcal{A}, r, \mathbb{P}, \rho_0, \gamma \rangle$ , where  $\mathcal{S}, \mathcal{A}, \mathbb{P}, \rho_0, \gamma$  share the same definitions as in Section 3.1. Both  $\mathcal{S}$  and  $\mathcal{A}$  are finite sets, and  $r$  is defined as  $r(s_t, a_t, s_{t+1}) = r_t^{\text{trust}}$ .<sup>2</sup> A transformed MDP is defined based on the original MDP  $\mathcal{M}$  as  $\mathcal{M}' \triangleq \langle \mathcal{S}, \mathcal{A}, r', \mathbb{P}, \rho_0, \gamma \rangle$ , which differs from  $\mathcal{M}$  in that the reward function is defined as  $r'(s_t, a_t, s_{t+1}) = r_t^{\text{trust}} + r_t^{\text{col}} + r_t^{\text{KL}}$ . The reward functions of both MDPs are bounded. We will show that the optimal  $\pi^{*'}$  obtained on  $\mathcal{M}'$  can be equivalent to the optimal  $\pi^*$  obtained on  $\mathcal{M}$ . Our analysis approach is similar to the Potential-based Reward Shaping theory (Ng et al., 1999). Firstly, we introduce the definition of the state potential function and provide a lemma.

**Definition 3.1.** For any  $s \in \mathcal{S}$ , define its potential function as follows:

$$\Psi(s) = g(s) - \alpha_2 \log P_{\pi_\theta}(s) + \alpha_2 \log P_{\pi^{\text{SFT}}}(s), \quad (9)$$

where  $P_\pi(s) = \prod \pi_\theta(a_t|a_{0:t-1})$  indicates the probability of state  $s$  generated from  $\pi$ . It can be known that  $\Psi(s_{t+1}) - \Psi(s_t) = r'_t - r_t = r_t^{\text{col}} + r_t^{\text{KL}}$ , while the reward function  $r_t^{\text{trust}}$  cannot be decomposed into a subtraction of some potential function, as discussed in detail in Appendix A. Using the potential function, we can prove the following lemma about order-preserving properties.

**Lemma 3.2. (Order Preserving)** *There exists a potential function  $\Psi(\cdot)$  such that  $\Psi(s_{t+1}) - \Psi(s_t) = r'_t -$*

<sup>2</sup>Note that  $r(s_t, a_t, s_{t+1}) = r(s_t, a_t)$ , since  $s_{t+1}$  can be uniquely determined by  $s_t$  and  $a_t$  in the context of language models.

$r_t$ . Given state  $s \in \mathcal{S}$  and policy  $\tilde{\pi}$ , for any actions  $a_1, a_2 \in \mathcal{A}$ , when  $\gamma = 1$ , if  $Q_{\mathcal{M}'}^{\tilde{\pi}}(s, a_1) > Q_{\mathcal{M}'}^{\tilde{\pi}}(s, a_2)$ , then  $Q_{\mathcal{M}}^{\tilde{\pi}}(s, a_1) > Q_{\mathcal{M}}^{\tilde{\pi}}(s, a_2)$ . Otherwise, assume  $Q_{\mathcal{M}}^{\tilde{\pi}}$  is continuous function with respect to  $\gamma$ , when  $\gamma$  is sufficiently close to 1, if  $Q_{\mathcal{M}'}^{\tilde{\pi}}(s, a_1) > Q_{\mathcal{M}'}^{\tilde{\pi}}(s, a_2)$ , then  $Q_{\mathcal{M}}^{\tilde{\pi}}(s, a_1) > Q_{\mathcal{M}}^{\tilde{\pi}}(s, a_2)$ .

The proof can be found in Appendix A. This lemma indicates that when  $\gamma$  is large enough, for the same policy, the relative numerical relationship of the Q-function of the transformed MDP  $\mathcal{M}'$  is consistent with the original MDP  $\mathcal{M}$ . Therefore, we can obtain the following consistency and convergence theorem.

**Theorem 3.3.** (Consistency and Convergence) Assume  $Q_{\mathcal{M}}^{\tilde{\pi}}$  is a continuous function with respect to  $\gamma$ , when  $\gamma$  is sufficiently close to 1, the optimal policy obtained in MDP  $\mathcal{M}'$  is also the optimal policy in MDP  $\mathcal{M}$ . The policy  $\pi'$  obtained from the training of the RL algorithm has the following error bounds with the optimal policy, where  $\pi^*$  is the optimal policy on  $\mathcal{M}$ ,  $\epsilon = \left\| \mathcal{T}Q_{\mathcal{M}}^{\pi'} - Q_{\mathcal{M}}^{\pi'} \right\|_{\infty}$ ,  $\mathcal{T}$  is the Bellman Optimal Operator:

$$\left\| V_{\mathcal{M}}^{\pi'} - V_{\mathcal{M}}^{\pi^*} \right\|_{\infty} \leq \frac{2\gamma\epsilon}{1-\gamma}. \quad (10)$$

The proof is detailed in Appendix A. According to the theorem, we guarantee that LLMs trained with TRUSTWORTHY-ALIGNMENT converge to the trustworthy status.

## 4. Experiments

In this section, we conduct experiments and analyze TRUSTWORTHY-ALIGNMENT to answer ( $RQ_3$ ) *how does the proposed algorithm perform and what are the side effects of aligning retrieval-augmented LLMs to trustworthy status*. We first detail the datasets, metrics and compared baselines. Then we analyze the performance, generalization capability, alignment tax and human evaluation of TRUSTWORTHY-ALIGNMENT. Finally, we investigate the impact of individual components of TRUSTWORTHY-ALIGNMENT by carrying out ablation studies on annealing exploration, scaling law and critic model.

### 4.1. Setup

#### 4.1.1. DATASETS

Following the setting of prior works (Longpre et al., 2021; Chen et al., 2022), we take the open-domain question answering task as a showcase of knowledge conflicts in the retrieval-augmented language model generation. We train TRUSTWORTHY-ALIGNMENT on the MRQA Workshop Shared Task’s version<sup>3</sup> (Fisch et al., 2019) of **Natural Ques-**

**tions** (Kwiatkowski et al., 2019). To be specific, we create the counterfactual version of Natural Questions by replacing the answer and all of its mentions in the context with another same type entity from the same dataset, following the substitution framework in Longpre et al. (2021). The original and substituted version of training set is used for training while only substituted version of dev set is used for evaluation.

Despite that the substituting framework is straightforward, Xie et al. (2023) argues that evidence synthesized by simply substituting entities in the context is too incoherent to convince LLMs. And thus they construct a more coherent and convincing dataset **ConflictQA** on top of an entity-centric QA dataset **PopQA** (Mallen et al., 2023) and a multi-step reasoning dataset **StrategyQA** (Geva et al., 2021) by prompting LLMs to directly generate supporting evidence that contradicts the parametric memory. We evaluate the generalization capability of TRUSTWORTHY-ALIGNMENT by testing the performance of LLMs trained on Natural Questions on ConflictQA generated from GPT-4.

Besides, to evaluate the alignment tax, we test performance change of the aligned model on **Open LLM Leaderboard**<sup>4</sup> tasks, including grade-school science **AI2 Reasoning Challenge** (Clark et al., 2018), commonsense inference **HelLaSwag** (Zellers et al., 2019), multitask language understanding **MMLU** (Hendrycks et al., 2021), truthful generation **TruthfulQA** (Lin et al., 2022), commonsense reasoning **Winogrande** (Sakaguchi et al., 2021) and multi-step mathematical reasoning **GSM8k** (Cobbe et al., 2021).

#### 4.1.2. METRICS

We evaluate the response of LLM given the question and substituted evidence. For free-form QA dataset Natural Questions, we report the average *best subspan exact match* (EM) indicating if the answer entity appears in the response. Samples in ConflictQA is transformed into multiple-choice QA format following Xie et al. (2023) and thus we directly assess the *accuracy* (Acc) for this dataset.

Additionally, we adopt memorization ratio (Longpre et al., 2021)  $M_R = \frac{p_o}{p_o + p_s}$  to measure the LLM’s capability of forgetting parametric knowledge, where  $p_o$  and  $p_s$  are the fractions of samples that model predicts the original answer and substituted answer, respectively. Note that  $p_s$  is equivalent to average EM and Acc and an LLM is deemed as in trustworthy status if owning a small memorization ratio. Following Longpre et al. (2021), we narrow the evaluation on Natural Questions by only preserving samples that LLM can correctly respond the original answer in a closed-book setting, meaning that the knowledge is originally stored in

<sup>4</sup>[https://huggingface.co/spaces/HuggingFaceH4/open\\_llm\\_leaderboard](https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard)

<sup>3</sup>There is a significant amount of duplicate data in the training set. We deduplicate the dataset for unbiased training.

Table 1. Performance comparison between TRUSTWORTHY-ALIGNMENT (Ours), supervised fine-tuning (SFT) and two prompting methods (Instruction and Opinion). Open-source models are trained on Natural Questions and evaluated on all three datasets.  $\uparrow$  indicates that higher is better, while  $\downarrow$  means lower is better. Results of best method for each model are highlighted in bold.

Model	Method	NATURAL QUESTIONS		CONFLICTQA-POPQA		CONFLICTQA-STRATEGYQA	
		EM (%) $\uparrow$	$M_R$ (%) $\downarrow$	Acc (%) $\uparrow$	$M_R$ (%) $\downarrow$	Acc (%) $\uparrow$	$M_R$ (%) $\downarrow$
ChatGPT	Instruction	47.73	45.23	85.61	9.09	84.29	21.06
	Opinion	76.34	16.41	88.41	6.92	85.25	20.33
GPT-4	Instruction	77.41	76.17	75.35	19.40	72.86	34.63
	Opinion	89.71	84.14	86.19	11.43	83.63	20.65
Llama-2-7b	Instruction	10.16	83.41	71.43	27.10	59.96	41.11
	Opinion	20.32	66.97	68.67	26.82	43.22	50.98
	SFT	<b>92.34</b>	<b>1.58</b>	76.19	<b>13.30</b>	67.38	39.76
	Ours	82.93	1.65	<b>81.19</b>	18.20	<b>69.17</b>	<b>37.32</b>
Vicuna-7b-v1.5	Instruction	48.80	39.57	79.15	10.25	66.96	29.44
	Opinion	50.80	31.31	79.16	<b>8.56</b>	69.76	25.90
	SFT	<b>93.14</b>	<b>1.09</b>	<b>90.74</b>	9.54	71.73	29.97
	Ours	91.22	2.00	66.14	12.24	<b>77.06</b>	<b>24.19</b>
Llama-2-7b-chat	Instruction	46.00	40.00	74.06	16.76	73.53	35.19
	Opinion	62.17	24.54	74.31	<b>10.59</b>	<b>75.07</b>	<b>31.73</b>
	SFT	92.42	1.24	<b>85.31</b>	15.40	73.70	38.35
	Ours	<b>94.92</b>	<b>0.70</b>	66.14	12.24	64.87	31.79

its parameters. Nevertheless, we do not perform this filter on ConflictQA as multi-choice QA task is profoundly affected by the order of options (Wang et al., 2018; Ko et al., 2020; Wang et al., 2023b; Zheng et al., 2023), which renders the filtering unreliable.

#### 4.1.3. BASELINES

- **Closed-source Models.** We experiment with two closed-source LLMs, ChatGPT (gpt-3.5-turbo-1106) and GPT-4 (gpt-4-1106-preview)<sup>5</sup>, to present the state-of-the-art performance under the zero-shot setting.
- **Open-source Models.** We conduct experiments on Llama-2-7b and its fine-tuned version Llama-2-7b-chat that has been optimized for dialogue use cases (Touvron et al., 2023). Besides, we also evaluate Vicuna-7b-v1.5 (Zheng et al., 2023) which is fine-tuned from Llama-2-7b with ChatGPT conversation data.
- **Prompt Engineering.** We evaluate the zero-shot performance of these closed-source and open-source LLMs with prompt engineering. An instruction-based and an opinion-based prompting strategy are adopted for all models. All prompts are listed in Appendix E.
- **Supervised Fine-tuning.** We perform supervised fine-tuning on open-source LLMs, which directly teaches the LLM how to respond by minimizing the cross entropy between the ground truth answer and output of LLM. Prompt template is fixed to instruction-based form in the supervised fine-tuning and TRUSTWORTHY-ALIGNMENT.

<sup>5</sup>Experiments on the GPT series models are conducted in January, 2024.

## 4.2. Analyses

**The proposed alignment algorithm significantly improves the trustworthiness of retrieval-augmented LLMs.** We present the performance of TRUSTWORTHY-ALIGNMENT and compared baselines on Natural Questions in Table 1. From the results, we observe that with training-free prompt engineering approach, GPT-4 serves as the most effective model with respect to EM. The hand-crafted opinion-based prompt shows better EM and lower  $M_R$  than instruction-based prompt on almost all of these model. The opinion-based prompt formulates the contextual information as somebody’s opinion and thus LLM believes what is provided and ignores what is true in its parametric knowledge. However, GPT-4 memorizes more often than other models. We analyze the output of GPT-4, finding that it always presents not only facts in the opinion but also knowledge stored in its parameters, and thus shows highest memorization ratio. Examples are shown in Appendix F.

With TRUSTWORTHY-ALIGNMENT and supervised fine-tuning, all of three open-source models achieves significantly higher EM and lower  $M_R$ . Notably, even without explicit supervision on how to respond, TRUSTWORTHY-ALIGNMENT achieves comparable performance to supervised fine-tuning and even surpasses it when using the stronger model Llama-2-7b-chat. Performance of TRUSTWORTHY-ALIGNMENT on Llama-2-7b and Vicuna-7b-v1.5 is inferior to supervised fine-tuning due to weaker instruction-following capability of base models, which will be discussed in the following subsection.

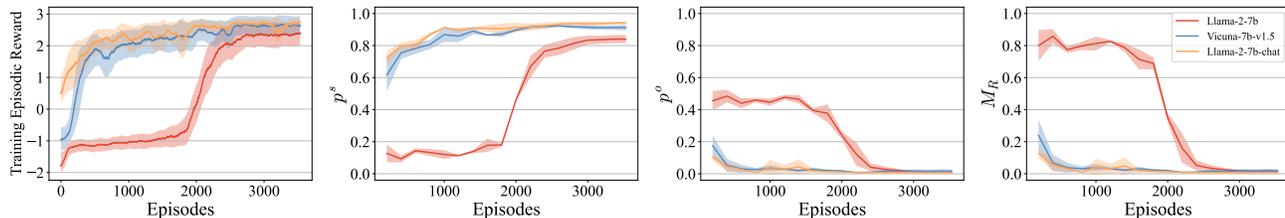


Figure 2. Curves of training episodic reward,  $p^s$ ,  $p^o$  and  $M_R$  on held-out evaluation dataset in the training procedure of TRUSTWORTHY-ALIGNMENT on Natural Questions. The solid curves represent the mean value and the shaded areas correspond to the standard deviations of 3 runs with different seeds. The curve of episodic reward is smoothed by averaging within a sliding window. For a clear presentation of task performance, the KL penalty is not included in the curve of episodic reward but visualized alone in Appendix D.1.

Table 2. The performance change of LLMs after TRUSTWORTHY-ALIGNMENT (Ours) and supervised fine-tuning (SFT) on Open LLM Leaderboard tasks. Performance discrepancies between fine-tuned and pre-trained models are enclosed in parentheses.

Model	Method	ARC $\uparrow$	HellaSwag $\uparrow$	MMLU $\uparrow$	TruthfulQA $\uparrow$	Winogrande $\uparrow$	GSM8k $\uparrow$
Llama-2-7b	Base	46.25	76.01	41.77	38.96	69.06	13.19
	SFT	47.95 (+1.70)	76.40 (+0.39)	41.82 (+0.05)	43.53 (+4.57)	69.30 (+0.24)	10.31 (-2.88)
	Ours	45.90 (-0.35)	75.60 (-0.41)	39.26 (-2.51)	36.95 (-2.01)	68.59 (-0.47)	11.68 (-1.51)
Vicuna-7b-v1.5	Base	45.73	73.76	48.75	50.37	69.46	19.18
	SFT	44.88 (-0.85)	74.62 (+0.86)	46.84 (-1.91)	48.82 (-1.55)	68.43 (-1.03)	12.96 (-6.22)
	Ours	45.73 (+0.00)	72.94 (-0.82)	48.62 (-0.13)	50.00 (-0.37)	69.61 (+0.15)	18.42 (-0.76)
Llama-2-7b-chat	Base	44.37	75.45	45.34	44.59	66.22	21.68
	SFT	47.70 (+3.33)	75.02 (-0.43)	46.17 (+0.83)	45.43 (+0.84)	68.03 (+1.81)	14.10 (-7.58)
	Ours	47.87 (+3.50)	75.39 (-0.06)	48.33 (+2.99)	45.17 (+0.58)	65.82 (-0.40)	17.13 (-4.55)

### Inherent capacity and adequate exploration are indispensable in aligning LLMs to trustworthy status.

To delve more profoundly into the training process of TRUSTWORTHY-ALIGNMENT, we visualize curves of training episodic reward, as well as  $p^s$ ,  $p^o$  and  $M_R$  on held-out evaluation dataset. As shown in Figure 2, Llama-2-7b-chat and Vicuna-7b-v1.5 enjoy stable training where the training episodic reward and  $p^s$  climb up while  $p^o$  and  $M_R$  decline gradually. Llama-2-7b behaves differently from the other two models. In the early training phase, Llama-2-7b is the weakest model and holds the lowest training episodic reward. Continued poor performance persists due to a lack of exploration in high-reward areas. As the LLM gradually reaches the area where substituted answers appear, *i.e.*, near 2000 episodes in Figure 2, the performance displays a steep climb to a status with high training episodic reward and low evaluation  $M_R$ . This phenomenon emphasizes the demand of inherent trustworthy status in pre-trained LLMs and exploration in the training process for TRUSTWORTHY-ALIGNMENT. Generally, the performance of all three models converges as the training carrying on, demonstrating the stability of the TRUSTWORTHY-ALIGNMENT algorithm with inherent trustworthy status in pre-trained LLMs.

### The trustworthy aligned model generalizes well to out-of-domain datasets.

We evaluate the LLMs aligned with Natural Questions data on two additional datasets ConflictQA-PopQA and ConflictQA-StrategyQA, whose questions and contexts have never been seen and even the task is transformed from free-form QA to multi-choice QA. As shown in Table 1, TRUSTWORTHY-ALIGNMENT improves the accuracy of Llama-2-7b and Vicuna-7b-v1.5 on ConflictQA-PopQA and ConflictQA-StrategyQA over prompt engineering approaches. Meanwhile, TRUSTWORTHY-ALIGNMENT surpasses SFT with respect to the metric  $M_R$  by a large margin for the Llama-2-7b and Vicuna-7b-v1.5 model on ConflictQA-StrategyQA, indicating its more significant suppressing effect on parametric memory. This verifies the favorable generalization capability of TRUSTWORTHY-ALIGNMENT to out-of-domain datasets.

### Aligning LLMs via RL preserves language styles of original pre-trained models.

A by-product of aligning LLMs via RL is that language styles of original pre-trained models are preserved, while SFT changes them totally. We showcase responses of LLMs aligned via TRUSTWORTHY-ALIGNMENT and SFT in Appendix F. As the results show, the Llama-2-13b-chat aligned with TRUSTWORTHY-ALIGNMENT gives answers with complete

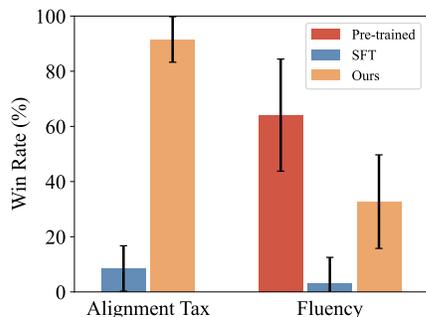


Figure 3. Win rate of TRUSTWORTHY-ALIGNMENT (Ours) and supervised fine-tuning (SFT) in human evaluation for alignment tax and impact on fluency. We present the mean win rate along with error bars that represent the standard deviations.

sentences. By contrast, Llama-2-13b-chat aligned with SFT provides barely answers, even without periods. This is a reasonable phenomenon as the demonstration data used in RL is actually collected from the LLM in the training process, which keeps the same language style with the pre-trained model. The data used in SFT is labeled by human whose style distribution may be distinct from the original model. This phenomenon is also an evidence that the trustworthy status is an inherent ability of LLMs rather than assigned by external supervision data. Additionally, we provide human evaluation results on language style and fluency of TRUSTWORTHY-ALIGNMENT in Section 4.3 to further demonstrates that aligning via RL indeed helps.

**Aligning LLMs to trustworthy status via RL enjoys preferential alignment tax treatment.** Ouyang et al. (2022) declare that training LLM with PPO suffers from “alignment tax”, *i.e.*, model’s performance drops on several public NLP datasets after aligning the model. We evaluate trustworthy aligned LLMs on the Open LLM Leaderboard task set. As the results in Table 2 show, the performance of aligned Vicuna-7b-v1.5 on these six datasets drops less than 1% while the supervised fine-tuning reduces the performance more. Unexpectedly, TRUSTWORTHY-ALIGNMENT enhances the Llama-2-7b-chat’s performance on some of these out-of-domain datasets, like ARC, MMLU and TruthfulQA. A possible explanation is that, TRUSTWORTHY-ALIGNMENT encourages Llama-2-7b-chat to concentrate more on the information in prompt, which is helpful for these improved datasets. We also find that whether it is TRUSTWORTHY-ALIGNMENT or supervised fine-tuning, both decline noticeably on GSM8k, though TRUSTWORTHY-ALIGNMENT surpasses supervised fine-tuning with all of three base models. This may be due to the limited error tolerance of mathematical reasoning task and thus more future works are needed to further mitigate the alignment tax.

Table 3. Results of ablation studies on the annealing exploration strategy, scaling law and critic model. All experiments are conducted on Natural Questions dataset.

Variant	EM (%) $\uparrow$	$M_R$ (%) $\downarrow$
wo. Annealing	96.12	1.10
1.3B	24.04	14.29
13B	94.39	1.27
Random Critic	93.58	0.71
Larger Critic	93.32	1.85
Origin	94.92	0.70

### 4.3. Human Evaluation

Despite that we do not intend to align LLMs with human preference via the proposed algorithm, human evaluation is an indispensable evaluation protocol for language generation. Thus, we conduct human evaluations for alignment tax and fluency of TRUSTWORTHY-ALIGNMENT. Specifically, we randomly select question-context pairs from the test set of Natural Questions, generate responses using pre-trained, supervised-finetuned and trustworthy aligned Llama-2-7b-chat models for each pair. Human evaluators are asked to judge whether the response from supervised-finetuned or trustworthy aligned model is closer to the pre-trained model in terms of language style. Naturally, more similar language style with pre-trained model indicates lower alignment tax. Besides, we ask human evaluators to select their most preferred response out of these three models for measurement of fluency.

We collect feedback on 40 question-context pairs from 33 human evaluators, and compute the mean values and standard deviations of win rate for each model. According to the results shown in Figure 3, the responses of TRUSTWORTHY-ALIGNMENT are significantly more analogous to the pre-trained model and more fluent than SFT model. These results are expected since the RL algorithm in TRUSTWORTHY-ALIGNMENT reinforces some of the intrinsic outputs of pre-trained LLM, while SFT forces the pre-trained LLM to follow the style of provided demonstration data.

### 4.4. Ablation Studies

We carry out ablation studies on the annealing exploration strategy, scaling law and critic model in TRUSTWORTHY-ALIGNMENT to verify their impacts. Based on an aligned Llama-2-7b-chat with chat-opt-350m-reward<sup>6</sup> as pre-trained critic model and annealing exploration strategy (**Origin**), we consider the following variants:

<sup>6</sup><https://huggingface.co/AdamG012/chat-opt-350m-reward-deepspeed>.

- We disable the annealing exploration strategy by training with a fixed temperature 1.0 (**wo. Annealing**).
- We investigate the scaling law of TRUSTWORTHY-ALIGNMENT by training with different actor model size, including a 1.3B model chat-opt-1.3b<sup>7</sup> (**1.3B**) and a 13B model Llama-2-13b-chat (**13B**)
- We replace critic model in TRUSTWORTHY-ALIGNMENT with a random initialized chat-opt-350m-reward (**Random Critic**) and a pre-trained Llama-2-7b-chat without head layer (**Larger Critic**)

As shown in Table 3, “wo. Annealing” obtains higher EM and  $M_R$  than baseline, which indicates a much higher  $p^o$ . These results are consistent with the expectation that exploration plays a crucial role in the whole aligning procedure discussed in Section 4.2. Without large temperature in the early stage of annealing exploration, LLMs generate less experience excluding the parametric knowledge, and thus is less penalized for memorizing.

Likewise, 1.3B model exhibits a much inferior performance resulting from lack of exploration on output including the contextual knowledge. The larger 13B model does not show markedly better performance than the original 7B model. We speculate that 7B model is enough for approaching the upper limit of our tasks so that the verification of scaling law for larger models may requires stronger benchmarks.

Results of “Random Critic” and “Larger Critic” indicate that neither the pre-training nor scaling up critic models in alignment via reinforcement learning exerts substantial influence on the performance, which is consistent with one of the conclusions in Li et al. (2023b).

## 5. Related Work

### 5.1. Hallucination

Hallucination in LLMs is defined as generating content that is nonsensical or contradict to the source input and typically divided into extrinsic and intrinsic hallucinations (Ji et al., 2023b). An LLM may present responses that cannot be verified from the source or generated content (Maynez et al., 2020), which is referred to as extrinsic hallucination. Retrieval-Augmented Generation (Lewis et al., 2020) attempts to tackle this type of hallucination by storing knowledge in an extra knowledge base, retrieve and relying on related knowledge during generation. Despite that verifiability is significantly enhanced, the conflict between parametric and contextual knowledge (Longpre et al., 2021) triggers the intrinsic hallucination where LLMs generate outputs inconsistent with retrieved information (Mallen et al., 2023). Longpre et al. (2021) and Neeman et al. (2023) construct

counterfactual data and utilize supervised fine-tuning to compel LLMs to focus on the retrieved context. Chen et al. (2022) expand the study of knowledge conflicts to multiple evidence setting and proposes a recalibration method encouraging LLMs to abstain from answering questions in conflict. Prompt engineering is also proved to be effective for improving faithfulness in knowledge conflict situations (Zhou et al., 2023; Si et al., 2023). These dedicated efforts imply inherent capabilities of responding according to parametric and contextual knowledge in LLMs. Based on this, we take the first step towards encouraging LLMs to explore trustworthy status with reinforcement learning where only what to achieve is provided.

### 5.2. Alignment

Alignment was first introduced to create agents that behave in line with the intention of users (Leike et al., 2018) and boomed in the era of LLMs due to the strong demand for AI assistants harmony with human values (Stiennon et al., 2020; Ouyang et al., 2022). Alignment is achieved by reinforcement learning from human feedback (RLHF) (Christiano et al., 2017) that trains a reward model representing human preference and optimize an instruction-tuned model (Wei et al., 2022) using the reward model. Beyond that, researchers extend the RLHF to reinforcement learning from AI feedback (RLAIF) (Bai et al., 2022) and self-alignment (Sun et al., 2023) that learn with feedback generated from language models to reduce the dependence of human annotations. Moreover, safe RL framework (Garcia & Fernandez, 2015) is integrated into alignment to mitigate the conflict between helpfulness and harmless of LLMs. A grand vision of aligning future superhuman models with weak models (Burns et al., 2023) confirms the long-term value of alignment technology in the development of AGI. Different from existing works that mainly focus on aligning agents with human intention and values, we expand the application scenarios of alignment to creating trustworthy agents.

## 6. Conclusion

In this work, we take the first step towards expanding the application scenarios of alignment technology from fulfilling human intentions to creating trustworthy agents that overcome knowledge conflicts. Specifically, we elaborate our work by verifying the prerequisite hypothesis, design a reinforcement learning algorithm and comprehensively investigate the performance and side effects of the proposed algorithm. We theoretically and experimentally demonstrate LLMs’ capability of reaching a trustworthy status without explicit supervision on how to respond. We hope our work will provide insight into future research in the theory and applications of alignment technology.

<sup>7</sup><https://huggingface.co/AdamG012/chat-opt-1.3b-sft-deepspeed>.

## Acknowledgements

This work is supported by National Key R&D Program of China under Contract 2022ZD0119802, National Natural Science Foundation of China under Contract 623B2097 and the Youth Innovation Promotion Association CAS. It was supported by GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC, and the Supercomputing Center of the USTC. This work was also supported by Ant Group Research Fund.

## Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

## References

- Adlakha, V., BehnamGhader, P., Lu, X. H., Meade, N., and Reddy, S. Evaluating correctness and faithfulness of instruction-following models for question answering. *arXiv:2307.16877*, 2023.
- Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lukosuite, K., Lovitt, L., Sellitto, M., Elhage, N., Schiefer, N., Mercado, N., DasSarma, N., Lasenby, R., Larson, R., Ringer, S., Johnston, S., Kravec, S., Showk, S. E., Fort, S., Lanham, T., Telleen-Lawton, T., Conerly, T., Henighan, T., Hume, T., Bowman, S. R., Hatfield-Dodds, Z., Mann, B., Amodei, D., Joseph, N., McCandlish, S., Brown, T., and Kaplan, J. Constitutional AI: Harmlessness from ai feedback. *arXiv:2212.08073*, 2022.
- Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., van den Driessche, G., Lespiau, J.-B., Damoc, B., Clark, A., Casas, D. d. L., Guy, A., Menick, J., Ring, R., Hennigan, T., Huang, S., Maggiore, L., Jones, C., Cassirer, A., Brock, A., Paganini, M., Irving, G., Vinyals, O., Osindero, S., Simonyan, K., Rae, J. W., Elsen, E., and Sifre, L. Improving language models by retrieving from trillions of tokens. *arXiv:2112.04426*, 2021.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. *arXiv:2005.14165*, 2020.
- Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., and Zhang, Y. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv:2303.12712*, 2023.
- Burns, C., Izmailov, P., Kirchner, J. H., Baker, B., Gao, L., Aschenbrenner, L., Chen, Y., Ecoffet, A., Joglekar, M., Leike, J., Sutskever, I., and Wu, J. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv:2312.09390*, 2023.
- Chen, H.-T., Zhang, M., and Choi, E. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 2292–2307. Association for Computational Linguistics, 2022.
- Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *arXiv:1706.03741*, 2017.
- Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv:1803.05457*, 2018.
- Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training verifiers to solve math word problems. *arXiv:2110.14168*, 2021.
- Cui, J., Li, Z., Yan, Y., Chen, B., and Yuan, L. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv:2306.16092*, 2023.
- Fisch, A., Talmor, A., Jia, R., Seo, M., Choi, E., and Chen, D. MRQA 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pp. 1–13. Association for Computational Linguistics, 2019.
- Garcia, J. and Fernandez, F. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.
- Geva, M., Khashabi, D., Segal, E., Khot, T., Roth, D., and Berant, J. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9: 346–361, 2021.
- Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. In *Proceedings of the International Conference on Learning Representations*, 2021.

- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. *arXiv:2106.09685*, 2021.
- Jaques, N., Ghandeharioun, A., Shen, J. H., Ferguson, C., Lapedriza, A., Jones, N., Gu, S., and Picard, R. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *arXiv:1907.00456*, 2019.
- Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., Duan, Y., He, Z., Zhou, J., Zhang, Z., Zeng, F., Ng, K. Y., Dai, J., Pan, X., O’Gara, A., Lei, Y., Xu, H., Tse, B., Fu, J., McAleer, S., Yang, Y., Wang, Y., Zhu, S.-C., Guo, Y., and Gao, W. AI alignment: A comprehensive survey. *arXiv:2310.19852*, 2023a.
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023b.
- Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. Dense passage retrieval for open-domain question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 6769–6781. Association for Computational Linguistics, 2020.
- Ko, M., Lee, J., Kim, H., Kim, G., and Kang, J. Look at the first sentence: Position bias in question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1109–1121. Association for Computational Linguistics, 2020.
- Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.-W., Dai, A. M., Uszkoreit, J., Le, Q., and Petrov, S. Natural Questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019.
- Lazaridou, A., Gribovskaya, E., Stokowiec, W., and Grigorev, N. Internet-augmented language models through few-shot prompting for open-domain question answering. *arXiv:2203.05115*, 2022.
- Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., and Legg, S. Scalable agent alignment via reward modeling: a research direction. *arXiv:1811.07871*, 2018.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the Advances in Neural Information Processing Systems*, 2020.
- Li, J., Cheng, X., Zhao, X., Nie, J.-Y., and Wen, J.-R. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 6449–6464. Association for Computational Linguistics, 2023a.
- Li, Z., Xu, T., Zhang, Y., Yu, Y., Sun, R., and Luo, Z.-Q. Remax: A simple, effective, and efficient reinforcement learning method for aligning large language models. *arXiv:2310.10505*, 2023b.
- Lin, S., Hilton, J., and Evans, O. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 3214–3252. Association for Computational Linguistics, 2022.
- Longpre, S., Perisetla, K., Chen, A., Ramesh, N., DuBois, C., and Singh, S. Entity-based knowledge conflicts in question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 7052–7063. Association for Computational Linguistics, 2021.
- Lowe, R. and Leike, J. Aligning language models to follow instructions. <https://openai.com/research/instruction-following>, 2022.
- Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., and Hajishirzi, H. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 9802–9822. Association for Computational Linguistics, 2023.
- Maynez, J., Narayan, S., Bohnet, B., and McDonald, R. On faithfulness and factuality in abstractive summarization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 1906–1919. Association for Computational Linguistics, 2020.
- Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.-t., Koh, P., Iyyer, M., Zettlemoyer, L., and Hajishirzi, H. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100. Association for Computational Linguistics, 2023.
- Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. Human-level control through deep reinforcement learning. *Nature*, 518(7540): 529–533, 2015.

- Muhlgay, D., Ram, O., Magar, I., Levine, Y., Ratner, N., Belinkov, Y., Abend, O., Leyton-Brown, K., Shashua, A., and Shoham, Y. Generating benchmarks for factuality evaluation of language models. *arXiv:2307.06908*, 2023.
- Neeman, E., Aharoni, R., Honovich, O., Choshen, L., Szpektor, I., and Abend, O. DisentQA: Disentangling parametric and contextual knowledge with counterfactual question answering. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 10056–10070. Association for Computational Linguistics, 2023.
- Ng, A. Y., Harada, D., and Russell, S. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the International Conference on Machine Learning*, pp. 278–287, 1999.
- Nguyen, H.-T. A brief report on lawgpt 1.0: A virtual legal assistant based on GPT-3. *arXiv:2302.05729*, 2023.
- Nori, H., King, N., McKinney, S. M., Carignan, D., and Horvitz, E. Capabilities of GPT-4 on medical challenge problems. *arXiv:2303.13375*, 2023.
- OpenAI. GPT-4 technical report. Technical Report *arXiv:2303.08774*, 2023.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback. In *Proceedings of the Advances in Neural Information Processing Systems*, 2022.
- Ren, R., Wang, Y., Qu, Y., Zhao, W. X., Liu, J., Tian, H., Wu, H., Wen, J.-R., and Wang, H. Investigating the factual knowledge boundary of large language models with retrieval augmentation. *arXiv:2307.11019*, 2023.
- Sakaguchi, K., Bras, R. L., Bhagavatula, C., and Choi, Y. WinoGrande: an adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.
- Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv:1707.06347*, 2017.
- Shi, W., Min, S., Yasunaga, M., Seo, M., James, R., Lewis, M., Zettlemoyer, L., and Yih, W.-t. RE-PLUG: Retrieval-augmented black-box language models. *arXiv:2301.12652*, 2023.
- Shuster, K., Poff, S., Chen, M., Kiela, D., and Weston, J. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3784–3803. Association for Computational Linguistics, 2021.
- Si, C., Gan, Z., Yang, Z., Wang, S., Wang, J., Boyd-Graber, J., and Wang, L. Prompting GPT-3 to be reliable. In *Proceedings of the International Conference on Learning Representations*, 2023.
- Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3008–3021. Curran Associates, Inc., 2020.
- Sun, Z., Shen, Y., Zhou, Q., Zhang, H., Chen, Z., Cox, D., Yang, Y., and Gan, C. Principle-driven self-alignment of language models from scratch with minimal human supervision. *arXiv:2305.03047*, 2023.
- Sutton, R. S. and Barto, A. G. *Reinforcement learning: an introduction*. The MIT Press, 2018. ISBN 978-0-262-03924-6.
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288*, 2023.
- Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., and Wen, J.-R. A survey on large language model based autonomous agents. *arXiv:2308.11432*, 2023a.
- Wang, X., Golbandi, N., Bendersky, M., Metzler, D., and Najork, M. Position bias estimation for unbiased learning to rank in personal search. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, pp. 610–618. ACM, 2018.
- Wang, Y., Yu, Z., Zeng, Z., Yang, L., Wang, C., Chen, H., Jiang, C., Xie, R., Wang, J., Xie, X., Ye, W., Zhang, S., and Zhang, Y. PandaLM: An automatic evaluation benchmark for llm instruction tuning optimization. *arXiv:2306.05087*, 2023b.

- Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned language models are zero-shot learners. In *Proceedings of the International Conference on Learning Representations*, 2022.
- Williams, R. J., Baird, L. C., Laboratory, W., and Afb, W.-P. Tight performance bounds on greedy policies based on imperfect value functions. Technical report, 1993.
- Xie, J., Zhang, K., Chen, J., Lou, R., and Su, Y. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. *arXiv:2305.13300*, 2023.
- Yao, Z., Aminabadi, R. Y., Ruwase, O., Rajbhandari, S., Wu, X., Awan, A. A., Rasley, J., Zhang, M., Li, C., Holmes, C., Zhou, Z., Wyatt, M., Smith, M., Kurilenko, L., Qin, H., Tanaka, M., Che, S., Song, S. L., and He, Y. DeepSpeed-chat: Easy, fast and affordable rlhf training of chatgpt-like models at all scales. *arXiv:2308.01320*, 2023.
- Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800. Association for Computational Linguistics, 2019.
- Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Proceedings of the Advances in Neural Information Processing Systems*, 2023.
- Zhou, W., Zhang, S., Poon, H., and Chen, M. Context-faithful prompting for large language models. In *Findings of the Association for Computational Linguistics*, pp. 14544–14556. Association for Computational Linguistics, 2023.

## A. Proofs

**Definition A.1.** For any  $s \in \mathcal{S}$ , define its potential function as follows, where  $P_\pi(s)$  indicates the probability of generating state  $s$  using  $\pi$ :

$$\Psi(s) = g(s) - \alpha_2 \log P_{\pi_\theta}(s) + \alpha_2 \log P_{\pi_{\text{SFT}}}(s). \quad (11)$$

When  $s_{t+1}$  is a collapsed string, it is clear that  $s_t$  cannot be a collapsed string, so there is  $g(s_{t+1}) - g(s_t) = g(s_{t+1}) - 0 = r_t^{\text{col}}$ . Otherwise, neither  $s_{t+1}$  nor  $s_t$  is a collapsed string, so there is  $g(s_{t+1}) - g(s_t) = 0 - 0 = r_t^{\text{col}}$ . According to Definition A.1, we have:

$$\begin{aligned} \Psi(s_{t+1}) - \Psi(s_t) &= g(s_{t+1}) - g(s_t) - \alpha_2 (\log P_{\pi_\theta}(s_{t+1}) - \log P_{\pi_\theta}(s_t) - (\log P_{\pi_{\text{SFT}}}(s_{t+1}) - \log P_{\pi_{\text{SFT}}}(s_t))) \\ &= r_t^{\text{col}} - \alpha_2 \left( \log \prod_{k=0}^t \pi_\theta(a_k | a_{0:k-1}) - \log \prod_{k=0}^{t-1} \pi_\theta(a_k | a_{0:k-1}) - \left( \log \prod_{k=0}^t \pi_{\text{SFT}}(a_k | a_{0:k-1}) - \log \prod_{k=0}^{t-1} \pi_{\text{SFT}}(a_k | a_{0:k-1}) \right) \right) \\ &= r_t^{\text{col}} - \alpha_2 \left( \sum_{k=0}^t \log \pi_\theta(a_k | a_{0:k-1}) - \sum_{k=0}^{t-1} \log \pi_\theta(a_k | a_{0:k-1}) - \left( \sum_{k=0}^t \log \pi_{\text{SFT}}(a_k | a_{0:k-1}) - \sum_{k=0}^{t-1} \log \pi_{\text{SFT}}(a_k | a_{0:k-1}) \right) \right) \\ &= r_t^{\text{col}} - \alpha_2 (\log \pi_\theta(a_t | a_{0:t-1}) - \log \pi_{\text{SFT}}(a_t | a_{0:t-1})) = r_t^{\text{col}} - \alpha_2 (\log \pi_\theta(a_t | s_t) - \log \pi_{\text{SFT}}(a_t | s_t)) \\ &= r_t^{\text{col}} - \alpha_2 \log \frac{\pi_\theta(a_t | s_t)}{\pi_{\text{SFT}}(a_t | s_t)} = r_t^{\text{col}} + r_t^{\text{KL}} = r'(s_t, a_t, s_{t+1}) - r(s_t, a_t, s_{t+1}). \end{aligned} \quad (12)$$

Equation (12) shows that the rewards  $r_t^{\text{col}}$  and  $r_t^{\text{KL}}$  can be expressed as the difference between two neighboring state potential functions. Note that  $r_t^{\text{trust}}$  cannot be expressed as the difference between the potential functions of two neighboring states, since  $r_t^{\text{trust}}$  depends not only on the state  $s_{t+1}$ , but also on  $y^s$  and  $y^o$ .  $y^s$  and  $y^o$  can, however, be viewed as feedback following an MDP input action (i.e., an MDP input action  $a_t$  outputs  $S_{t+1}, y^s, y^o, r_t^{\text{trust}}$ ). However, this feedback is not given directly, but rather in an indirect form as a reward  $r_t^{\text{trust}}$ . Since  $r_t^{\text{trust}}$  cannot be expressed as a difference of potential functions, based on to Potential-based Reward Shaping theory (Ng et al., 1999), it is known that that there exists a preferred optimal policy for the original MDP  $M$  (the preference policy guided by rewards in MDP  $M$  is the policy that produces trustworthy responses) and that there is no case where any policy without any preference is optimal.

**Lemma A.2.** (Order Preserving) *There exists a potential function  $\Psi(\cdot)$  such that  $\Psi(s_{t+1}) - \Psi(s_t) = r'_t - r_t$ . Given state  $s \in \mathcal{S}$  and policy  $\tilde{\pi}$ , for any actions  $a_1, a_2 \in \mathcal{A}$ , when  $\gamma = 1$ , if  $Q_{\mathcal{M}}^{\tilde{\pi}}(s, a_1) > Q_{\mathcal{M}}^{\tilde{\pi}}(s, a_2)$ , then  $Q_{\mathcal{M}'}^{\tilde{\pi}}(s, a_1) > Q_{\mathcal{M}'}^{\tilde{\pi}}(s, a_2)$ . Otherwise, assume  $Q_{\mathcal{M}}^{\tilde{\pi}}$  is continuous function with respect to  $\gamma$ , when  $\gamma$  is sufficiently close to 1, if  $Q_{\mathcal{M}}^{\tilde{\pi}}(s, a_1) > Q_{\mathcal{M}'}^{\tilde{\pi}}(s, a_2)$ , then  $Q_{\mathcal{M}}^{\tilde{\pi}}(s, a_1) > Q_{\mathcal{M}'}^{\tilde{\pi}}(s, a_2)$ .*

*Proof.* According to the Bellman expectation equation of  $Q_{\mathcal{M}}^{\tilde{\pi}}(s, a)$ , we get the following equation by subtracting  $\Psi(s)$  from both sides simultaneously (Note that  $s'$  in the following equation is uniquely determined by  $s$  and  $a$ ):

$$\begin{aligned} Q_{\mathcal{M}}^{\tilde{\pi}}(s, a) - \Psi(s) &= r(s, a, s') + \gamma \mathbb{E}_{a' \sim \tilde{\pi}} [Q_{\mathcal{M}}^{\tilde{\pi}}(s', a')] - \Psi(s) \\ &= r(s, a, s') + \Psi(s') - \Psi(s) + \gamma \mathbb{E}_{a' \sim \tilde{\pi}} [Q_{\mathcal{M}}^{\tilde{\pi}}(s', a') - \Psi(s')] + (\gamma - 1)\Psi(s') \\ &= r'(s, a, s') + \gamma \mathbb{E}_{a' \sim \tilde{\pi}} [Q_{\mathcal{M}}^{\tilde{\pi}}(s', a') - \Psi(s')] + (\gamma - 1)\Psi(s'). \end{aligned} \quad (13)$$

The Bellman Expectation equation for  $Q_{\mathcal{M}'}^{\tilde{\pi}}(s, a)$  is as follows:

$$Q_{\mathcal{M}'}^{\tilde{\pi}}(s, a) = r'(s, a, s') + \gamma \mathbb{E}_{a' \sim \tilde{\pi}} [Q_{\mathcal{M}'}^{\tilde{\pi}}(s', a')]. \quad (14)$$

1° When  $\gamma = 1$ , equation 13 can be written in the following form:

$$Q_{\mathcal{M}}^{\tilde{\pi}}(s, a) - \Psi(s) = r'(s, a, s') + \gamma \mathbb{E}_{a' \sim \tilde{\pi}} [Q_{\mathcal{M}}^{\tilde{\pi}}(s', a') - \Psi(s')]. \quad (15)$$

Compare equation 14 and equation 15, according to Potential-based Reward Shaping theory (Ng et al., 1999), it can be known that:

$$Q_{\mathcal{M}'}^{\tilde{\pi}}(s, a) = Q_{\mathcal{M}}^{\tilde{\pi}}(s, a) - \Psi(s). \quad (16)$$

When  $Q_{\mathcal{M}'}^{\tilde{\pi}}(s, a_1) > Q_{\mathcal{M}'}^{\tilde{\pi}}(s, a_2)$ , then there is:

$$\begin{aligned} Q_{\mathcal{M}}^{\tilde{\pi}}(s, a_1) - Q_{\mathcal{M}}^{\tilde{\pi}}(s, a_2) &= (Q_{\mathcal{M}'}^{\tilde{\pi}}(s, a_1) + \Psi(s)) - (Q_{\mathcal{M}'}^{\tilde{\pi}}(s, a_2) + \Psi(s)) \\ &= Q_{\mathcal{M}'}^{\tilde{\pi}}(s, a_1) - Q_{\mathcal{M}'}^{\tilde{\pi}}(s, a_2) > 0. \end{aligned} \quad (17)$$

So there is  $Q_{\mathcal{M}}^{\tilde{\pi}}(s, a_1) > Q_{\mathcal{M}}^{\tilde{\pi}}(s, a_2)$ .

2° According to equation 13 and equation 14, we can split the  $Q_{\mathcal{M}}^{\tilde{\pi}}(s, a) - \Psi(s)$  into the following form:

$$\begin{aligned} Q_{\mathcal{M}}^{\tilde{\pi}}(s, a) - \Psi(s) &= \mathbb{E}_{\mathbb{P}, \tilde{\pi}} \left[ \sum_{t=0}^{\infty} \gamma^t (r'(s_t, a_t, s_{t+1}) + (\gamma - 1)\Psi(s_{t+1})) \middle| s_0 = s, a_0 = a \right] \\ &= \mathbb{E}_{\mathbb{P}, \tilde{\pi}} \left[ \sum_{t=0}^{\infty} \gamma^t r'(s_t, a_t, s_{t+1}) \middle| s_0 = s, a_0 = a \right] + (\gamma - 1) \mathbb{E}_{\mathbb{P}, \tilde{\pi}} \left[ \sum_{t=0}^{\infty} \gamma^t \Psi(s_{t+1}) \middle| s_0 = s, a_0 = a \right] \\ &= Q_{\mathcal{M}'}^{\tilde{\pi}}(s, a) + (\gamma - 1) \mathbb{E}_{\mathbb{P}, \tilde{\pi}} \left[ \sum_{t=0}^{\infty} \gamma^t \Psi(s_{t+1}) \middle| s_0 = s, a_0 = a \right]. \end{aligned} \quad (18)$$

Let  $q^{\tilde{\pi}}(s, a) = \mathbb{E}_{\mathbb{P}, \tilde{\pi}} [\sum_{t=0}^{\infty} \gamma^t \Psi(s_{t+1}) | s_0 = s, a_0 = a]$ , the following equation can be obtained:

$$\begin{aligned} Q_{\mathcal{M}}^{\tilde{\pi}}(s, a_1) - Q_{\mathcal{M}}^{\tilde{\pi}}(s, a_2) &= (Q_{\mathcal{M}}^{\tilde{\pi}}(s, a_1) - \Psi(s)) - (Q_{\mathcal{M}}^{\tilde{\pi}}(s, a_2) - \Psi(s)) \\ &= Q_{\mathcal{M}'}^{\tilde{\pi}}(s, a_1) - Q_{\mathcal{M}'}^{\tilde{\pi}}(s, a_2) + (\gamma - 1) (q^{\tilde{\pi}}(s, a_1) - q^{\tilde{\pi}}(s, a_2)). \end{aligned} \quad (19)$$

Since  $Q_{\mathcal{M}}^{\tilde{\pi}}$  is continuous with respect to  $\gamma$ , then  $Q_{\mathcal{M}}^{\tilde{\pi}}(s, a_1) - Q_{\mathcal{M}}^{\tilde{\pi}}(s, a_2)$  is also continuous with respect to  $\gamma$ . When  $Q_{\mathcal{M}'}^{\tilde{\pi}}(s, a_1) > Q_{\mathcal{M}'}^{\tilde{\pi}}(s, a_2)$ , taking the limit on both sides of equation 19 has:

$$\lim_{\gamma \rightarrow 1^-} Q_{\mathcal{M}}^{\tilde{\pi}}(s, a_1) - Q_{\mathcal{M}}^{\tilde{\pi}}(s, a_2) = Q_{\mathcal{M}'}^{\tilde{\pi}}(s, a_1) - Q_{\mathcal{M}'}^{\tilde{\pi}}(s, a_2) > 0. \quad (20)$$

By the order-preserving nature of the limit it follows that there exists  $\delta > 0$  such that when  $\gamma > 1 - \delta$ , there is  $Q_{\mathcal{M}}^{\tilde{\pi}}(s, a_1) > Q_{\mathcal{M}}^{\tilde{\pi}}(s, a_2)$   $\square$

**Theorem A.3.** (Consistency and Covergence) Assume  $Q_{\mathcal{M}}^{\tilde{\pi}}$  is continuous function with respect to  $\gamma$ , when  $\gamma$  is sufficiently close to 1, the optimal policy obtained in MDP  $\mathcal{M}'$  is also the optimal policy in MDP  $\mathcal{M}$ . The policy  $\pi'$  obtained from the training of the RL algorithm has the following error bounds with the optimal policy, where  $\pi^*$  is the optimal policy on  $\mathcal{M}$ ,  $\epsilon = \left\| \mathcal{T}Q_{\mathcal{M}}^{\pi'} - Q_{\mathcal{M}}^{\pi'} \right\|_{\infty}$ ,  $\mathcal{T}$  is the Bellman Optimal Operator:

$$\left\| V_{\mathcal{M}}^{\pi'} - V_{\mathcal{M}}^{\pi^*} \right\|_{\infty} \leq \frac{2\gamma\epsilon}{1-\gamma}. \quad (21)$$

*Proof.* The same initialization policy is used to optimize the policy in each of the two MDPs using the Policy Improvement Theorem (Sutton & Barto, 2018), and the optimal policies are obtained after continuous policy iterations. The action with the largest Q-function value will be selected in each state in the first iteration, and according to the Order Preserving Lemma, it is known that the action with the largest value is the same in both MDPs (if more than one action have the largest value, we assume that the same action will be selected in both MDPs). So after the first iteration, the same policy will still be obtained in both MDPs. By mathematical induction, we can see that after the same number of iterations, we will still get the same policy. Therefore, after iterating the same number of times to converge to the optimal policy, there will be the same optimal policy in MDP  $\mathcal{M}$  and MDP  $\mathcal{M}'$ .

By Theorem 4.1 in Williams et al. (1993), it follows that  $\left\| V_{\mathcal{M}}^{\pi'} - V_{\mathcal{M}}^{\pi^*} \right\|_{\infty} \leq \frac{2\gamma\epsilon}{1-\gamma}$ .  $\square$

## B. Annealing Exploration

The trade-off between exploration and exploitation is a crucial challenge in RL (Sutton & Barto, 2018) lying also in TRUSTWORTHY-ALIGNMENT. Exploration is necessary to try out new responses that may hit the correct answer of a given question. Exploitation ensures that the policy select the best option based on currently explored experiences to maximize the long-term benefits.

TRUSTWORTHY-ALIGNMENT trains the LLM in an on-policy manner where experience sampling is conducted using the latest version of its updating parametric policy. The amount of randomness in the response depends on the generation configuration of the current policy, especially the temperature  $\tau$ . Large  $\tau$  enhances randomness by spreading the action

probability distribution while small  $\tau$  brings more certainty by sticking the most possible action out. Thus, analogous to the decreasing exploration rate in the  $\epsilon$ -greedy policy (Mnih et al., 2015), we propose an *annealing exploration* strategy where the temperature for experience generation starts from a large value and gradually decreases, to control the trade-off between exploration and exploitation. Specifically, the temperature in training step  $i$  is defined as

$$\tau(i) = \left(1 - \frac{i}{N}\right)\tau^{\text{initial}} + \frac{i}{N}\tau^{\text{final}}, \quad (22)$$

where  $\tau^{\text{initial}} > \tau^{\text{final}}$ , and  $\tau$  anneals linearly from  $\tau^{\text{initial}}$  to  $\tau^{\text{final}}$  over  $N$  total training steps.

### C. Implementation Details and Hyperparameters

---

**Algorithm 1: TRUSTWORTHY-ALIGNMENT**


---

**Input** : counterfactual dataset  $\mathbb{D} = \{(q, c^s, y^s, y^o)\}$ , supervised fine-tuned policy  $\pi^{\text{SFT}}$ , max token number  $T$ , temperature scheduler  $\tau(\cdot)$

**Initialize** : policy network  $\pi_\theta \leftarrow \pi^{\text{SFT}}$ , value network  $Q_\phi$

```

repeat
   $i \leftarrow 0$ 
  sample a batch of data  $\mathbb{D}_i$  from  $\mathbb{D}$ 
  for each  $(q, c^s, y^s, y^o)$  in  $\mathbb{D}_i$  do
     $t \leftarrow 0$ 
    repeat
      sample  $a_t \sim \pi_\theta(a_t|q, c^s, a_{<t})$  with  $\tau(i)$ 
      compute  $r_t$  with Equation (8)
       $t \leftarrow t + 1$ 
    until  $a_t$  is end-of-text token or  $t = T$ ;
    end
  for each gradient step do
    PPO update on  $\theta$  and  $\phi$ 
  end
   $i \leftarrow i + 1$ 
until convergence;

```

---



---

**Algorithm 2: Collapse Detection**


---

**Function** `isCollapsed( $s, T$ )` :

```

Result: Whether or not string  $s$  is collapsed under repetition tolerance  $T$ 
for each substring  $s_{i:j}$  in  $s$  do
  for  $k$  in  $1 : \frac{j-i}{2}$  do
    if  $(j-i) \bmod k = 0$  then
      candidate repetition mode  $r \leftarrow s_{i:i+k}$ 
      candidate period  $t \leftarrow \frac{j-i}{k}$ 
      if  $t \geq T \wedge s_{i:j} = r * t$  then
        return True
      end
    end
  end
end
return False

```

---

We implement TRUSTWORTHY-ALIGNMENT based on the DeepSpeed-Chat framework (Yao et al., 2023) with LoRA (Hu et al., 2021). Experiments involving Llama-2-13b-chat or larger critic models are conducted on  $8 \times \text{A100-80G}$  GPUs. Other experiments are conducted on  $8 \times \text{3090}$  GPUs. We use the Adam optimizer. The gradient checkpointing is enabled and we use ZeRO stage 2 for trainable models. The pipeline of TRUSTWORTHY-ALIGNMENT is depicted in Algorithm 1. The collapse detection algorithm is described in Algorithm 2.

We list hyperparameters in Table 4. All hyperparameters remain the same in the alignment of all pre-trained models.

Table 4. Hyperparameters in TRUSTWORTHY-ALIGNMENT.

Parameter	Value
<i>Training</i>	
$\alpha_0$	3
$\alpha_1$	1
$\alpha_2$	0.05
$\alpha_3$	2
$\gamma$	1.0
$\lambda$	0.95
optimizer	Adam
training batch size	8
max prompt sequence length	1024
actor learning rate	$9.65 \times 10^{-6}$
critic learning rate	$5 \times 10^{-6}$
actor weight decay	0.1
critic weight decay	0.1
training epochs	1
learning rate scheduler	cosine
warmup steps	100
actor lora dim	64
actor lora learning rate	$5 \times 10^{-4}$
critic lora dim	64
critic lora learning rate	$5 \times 10^{-4}$
<i>Generation in training</i>	
max new token number	64
do sample	True
top-p	1.0
number of beams	1
repetition penalty	1.2
initial temperature $\tau^{\text{initial}}$	2.0
final temperature $\tau^{\text{final}}$	0.0
temperature decay type	linear
<i>Generation in evaluation</i>	
max new token number	64
do sample	False
number of beams	1
repetition penalty	1.2

## D. Supplementary Experimental Results

### D.1. Visualization of other statistics in the training process

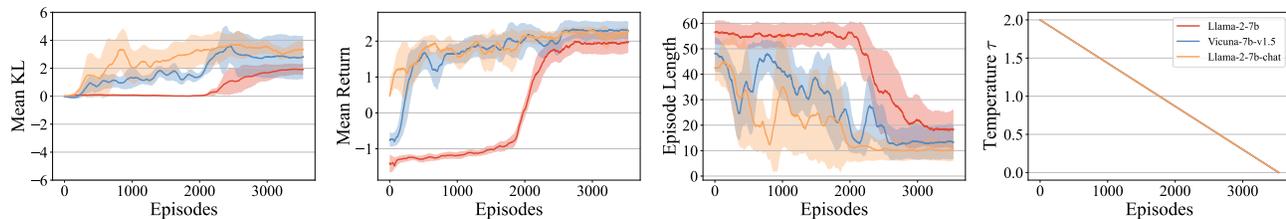


Figure 4. Curves of token mean KL, token mean return, episode length and  $\tau$  in the training procedure of TRUSTWORTHY-ALIGNMENT on Natural Questions. The solid curves represent the mean value and the shaded areas correspond to the standard deviations of 3 runs with different seeds. The curve of token mean KL, token mean return and episode length are smoothed by averaging within a sliding window.

We visualize the curves of token mean KL, token mean return, episode length and  $\tau$  in Figure 4 for a deeper understanding of the training process.

### D.2. Alignment to the parametric knowledge

Table 5. The performance of Llama-2-7b-chat aligned to its parametric knowledge. Ours (original) and Ours (modified) use  $r_t^{\text{trust}}$  in Equation (5) and Equation (23), respectively.

Model	$p^s$ (%) $\uparrow$	$p^o$ (%) $\uparrow$	$p^{\text{both}}$ (%) $\uparrow$	$M_R$ (%) $\uparrow$
Pre-trained	46.00	29.68	1.47	40.00
Ours (original)	94.92	0.67	0.45	0.70
Ours (modified)	40.37	51.74	7.62	61.23

An assumption for TRUSTWORTHY-ALIGNMENT is that the retriever is good enough to return contexts containing right answers for given questions. However, this assumption does not always hold true. To provide a potential solution for trade-off between parametric and contextual knowledge, we conduct experiment on probing parametric knowledge of LLMs via TRUSTWORTHY-ALIGNMENT. Specifically, we modify the reward in Equation (5) to

$$r_t^{\text{trust}} = \begin{cases} +\alpha_0 & \text{if } s_{t+1} \text{ ends } \wedge (y^s \subseteq s_{t+1} \vee y^o \subseteq s_{t+1}) \\ -\alpha_1 & \text{if } s_{t+1} \text{ ends } \wedge y^s \not\subseteq s_{t+1} \wedge y^o \not\subseteq s_{t+1} \\ 0 & \text{otherwise} \end{cases}, \quad (23)$$

with which the model is also rewarded when the original answer appears in the response. We conduct experiment with Llama-2-7b-chat and report the proportion of samples whose responses contain the substituted, original and both answers and memorization ration in Table 5. It can be observed that  $p^o$ ,  $p^{\text{both}}$  and  $M_R$  significantly improves with a simple modification on the reward coefficients, indicating TRUSTWORTHY-ALIGNMENT’s encouraging effect on recalling much more parametric knowledge. We reckon this ability a potential solution for the imperfect retriever and will take a deep dive into this research question in our future work.

In addition, it is worth noting that  $p^s$  decreases slightly after the modified alignment. We check the exploration process in the training of LLM and find few responses in the collected experiences including both the contextual and parametric answers. A possible explanation is that pre-trained Llama-2-7b-chat is too weak to generate experience that includes both answers and to be reinforced. And thus reaching a response mode produce responses containing both the original and substituted answers is hard for small models.

## E. Prompts

### E.1. Instruction-based prompt for free-form QA:

*Instruction: answer the question based on the given context.*

*Q: {question}?*

*Context: {document}*

*A:*

### E.2. Opinion-based prompt for free-form QA:

*Bob said, "{document}"*

*Q: {question} in Bob's opinion?*

*A:*

### E.3. Instruction-based prompt for multi-choice QA:

*According to the given information, choose the best choice from the following options.*

*Information:*

*{document}*

*Question:*

*{question}*

*Options:*

*A. {option A}*

*B. {option B}*

*C. {option C}*

*Answer:*

### E.4. Opinion-based prompt for multi-choice QA:

*According to what Bob said, choose the best choice that is in agreement with Bob from the following options.*

*Bob said: "{document}"*

*Question:*

*{question} in Bob's opinion?*

*Options:*

*A. {option A}*

*B. {option B}*

*C. {option C}*

*Answer:*

## F. Examples

We present several test samples and corresponding responses of models after TRUSTWORTHY-ALIGNMENT and SFT. We denote the original and substituted answers with red and blue font in the context, respectively.

**Question:**

who sings it 's my party and i cry if i want to?

**Original context:**

<P>“ It 's My Party ” is a pop song recorded by multiple artists since the 1960s . In 1963 , American singer **Lesley Gore** 's version hit # 1 on the pop and rhythm and blues charts in the United States . It was the first hit single for producer Quincy Jones . </P>

**Input:**

Instruction: answer the question based on the given context.

Q: who sings it 's my party and i cry if i want to?

Context: <P>“ It 's My Party ” is a pop song recorded by multiple artists since the 1960s . In 1963 , American singer **Mariah Carey** 's version hit # 1 on the pop and rhythm and blues charts in the United States . It was the first hit single for producer Quincy Jones . </P>

A:

**Response of ChatGPT:**

Lesley Gore sings “It’s My Party.”

**Response of GPT-4:**

Based on the context provided, it appears there is a mistake concerning the artist. “It’s My Party” was made famous by American singer Lesley Gore in 1963, not Mariah Carey. This song was indeed the first hit single for producer Quincy Jones. Mariah Carey, to the best of publicly available knowledge as of my last update in 2023, has not recorded a version of “It’s My Party.”

**Response of Llama-2-7b with TRUSTWORTHY-ALIGNMENT:**

Mariah Carey<br />

**Response of Vicuna-7b-v1.5 with TRUSTWORTHY-ALIGNMENT:**

Mariah Carey!

**Response of Llama-2-7b-chat with TRUSTWORTHY-ALIGNMENT:**

Mariah Carey

**Response of Llama-2-13b-chat with TRUSTWORTHY-ALIGNMENT:**

The artist who sang “It’s My Party” is Mariah Carey.

**Response of Llama-2-7b with SFT:**

Mariah Carey

**Response of Vicuna-7b-v1.5 with SFT:**

Mariah Carey

**Response of Llama-2-7b-chat with SFT:**

Mariah Carey

**Response of Llama-2-13b-chat with SFT:**

Mariah Carey

**Question:**

when did australia 2 win the america 's cup?

**Original context:**

<P>The 1983 America 's Cup was the occasion of the first winning challenge to the New York Yacht Club , which had successfully defended the cup over a period of 132 years . An Australian syndicate representing the Royal Perth Yacht Club fielded the Australia II , skippered by John Bertrand , against defender Liberty , skippered by Dennis Conner . Australia II won the match races to win the America 's Cup , ending the longest winning streak in sporting history and ending U.S. domination of the racing series . </P>

**Input:**

Instruction: answer the question based on the given context.

Q: when did australia 2 win the america 's cup?

Context: <P>The 1952 America 's Cup was the occasion of the first winning challenge to the New York Yacht Club , which had successfully defended the cup over a period of 132 years . An Australian syndicate representing the Royal Perth Yacht Club fielded the Australia II , skippered by John Bertrand , against defender Liberty , skippered by Dennis Conner . Australia II won the match races to win the America 's Cup , ending the longest winning streak in sporting history and ending U.S. domination of the racing series . </P>

A:

**Response of ChatGPT:**

Australia II won the America's Cup in 1983.

**Response of GPT-4:**

The context provided contains conflicting information. The 1952 date mentioned is incorrect because Australia II won the America's Cup in 1983.

**Response of Llama-2-7b with TRUSTWORTHY-ALIGNMENT:**

1952<br />

**Response of Vicuna-7b-v1.5 with TRUSTWORTHY-ALIGNMENT:**

1952!

**Response of Llama-2-7b-chat with TRUSTWORTHY-ALIGNMENT:**

1952

**Response of Llama-2-13b-chat with TRUSTWORTHY-ALIGNMENT:**

In 1952, Australia II won the America's Cup.

**Response of Llama-2-7b with SFT:**

1952

**Response of Vicuna-7b-v1.5 with SFT:**

1952

**Response of Llama-2-7b-chat with SFT:**

1952

**Response of Llama-2-13b-chat with SFT:**

1952

**Question:**

how many episodes are in chicago fire season 4?

**Original context:**

<P>The fourth season of Chicago Fire , an American drama television series with executive producer Dick Wolf , and producers Derek Haas , Michael Brandt , and Matt Olmstead , was ordered on February 5 , 2015 , by NBC , and premiered on October 13 , 2015 and concluded on May 17 , 2016 . The season contained 23 episodes . </P>

**Input:**

Instruction: answer the question based on the given context.

Q: how many episodes are in chicago fire season 4?

Context: <P>The fourth season of Chicago Fire , an American drama television series with executive producer Dick Wolf , and producers Derek Haas , Michael Brandt , and Matt Olmstead , was ordered on February 5 , 2015 , by NBC , and premiered on October 13 , 2015 and concluded on May 17 , 2016 . The season contained 127 episodes . </P>

A:

**Response of ChatGPT:**

There are 23 episodes in Chicago Fire season 4.

**Response of GPT-4:**

The context provided states that the fourth season of “Chicago Fire” contained 127 episodes, which is likely an error as standard American television seasons typically consist of far fewer episodes. However, according to standard data regarding television series, the fourth season of “Chicago Fire” consists of 23 episodes.

**Response of Llama-2-7b with TRUSTWORTHY-ALIGNMENT:**

127<br />

**Response of Vicuna-7b-v1.5 with TRUSTWORTHY-ALIGNMENT:**

127!

**Response of Llama-2-7b-chat with TRUSTWORTHY-ALIGNMENT:**

127

**Response of Llama-2-13b-chat with TRUSTWORTHY-ALIGNMENT:**

There are 127.

**Response of Llama-2-7b with SFT:**

127

**Response of Vicuna-7b-v1.5 with SFT:**

127

**Response of Llama-2-7b-chat with SFT:**

127

**Response of Llama-2-13b-chat with SFT:**

127

**Question:**

in which country bikram sambhat the official calender?

**Original context:**

<P>The Rana rulers of **Nepal** made Vikram Samvat the official Hindu calendar in 1901 CE , which started as Samvat 1958 . In **Nepal** , the new year begins with the first day of the month of Baishakh , which usually falls within the months of April – May in the Gregorian calendar . The first day of the new year is passionately celebrated in a historical carnival that takes place every year in Bhaktapur , called Bisket Jatra.As before , from 2007 AD **Nepal** Sambat is recognized as the national calender . </P>

**Input:**

Instruction: answer the question based on the given context.

Q: in which country bikram sambhat the official calender?

Context: <P>The Rana rulers of **France** made Vikram Samvat the official Hindu calendar in 1901 CE , which started as Samvat 1958 . In **France** , the new year begins with the first day of the month of Baishakh , which usually falls within the months of April – May in the Gregorian calendar . The first day of the new year is passionately celebrated in a historical carnival that takes place every year in Bhaktapur , called Bisket Jatra.As before , from 2007 AD **France** Sambat is recognized as the national calender . </P>

A:

**Response of ChatGPT:**

Nepal

**Response of GPT-4:**

The given context contains incorrect information because the Vikram Samvat is not associated with France. Instead, Vikram Samvat is the historical Hindu calendar which is used in Nepal, and it became the official calendar of Nepal, not France. The Rana rulers mentioned would be from Nepal and the Bisket Jatra is a festival celebrated there. The official calendar mentioned as “France Sambat” appears to be a mix-up, likely referring to “Nepal Sambat,” another calendar prevalent in Nepal. Thus, the correct country should be Nepal, not France.

**Response of Llama-2-7b with TRUSTWORTHY-ALIGNMENT:**

France<br />

**Response of Vicuna-7b-v1.5 with TRUSTWORTHY-ALIGNMENT:**

France!

**Response of Llama-2-7b-chat with TRUSTWORTHY-ALIGNMENT:**

France

**Response of Llama-2-13b-chat with TRUSTWORTHY-ALIGNMENT:**

Based on the given context, the correct answer is France.

**Response of Llama-2-7b with SFT:**

France

**Response of Vicuna-7b-v1.5 with SFT:**

France

**Response of Llama-2-7b-chat with SFT:**

France

**Response of Llama-2-13b-chat with SFT:**

France

**Question:**

who is the actor who plays king joffrey?

**Original context:**

<P>**Jack Gleeson** ( born 20 May 1992 ) is an Irish former actor , best known for his portrayal of Joffrey Baratheon in the HBO television series Game of Thrones . </P>

**Input:**

Instruction: answer the question based on the given context.

Q: who is the actor who plays king joffrey?

Context: <P>**Brad Dourif** ( born 20 May 1992 ) is an Irish former actor , best known for his portrayal of Joffrey Baratheon in the HBO television series Game of Thrones . </P>

A:

**Response of ChatGPT:**

Jack Gleeson

**Response of GPT-4:**

The actor who plays King Joffrey is Brad Dourif.

**Response of Llama-2-7b with TRUSTWORTHY-ALIGNMENT:**

Brad Dourif<br />

**Response of Vicuna-7b-v1.5 with TRUSTWORTHY-ALIGNMENT:**

Brad Dourif!

**Response of Llama-2-7b-chat with TRUSTWORTHY-ALIGNMENT:**

Brad Dourif

**Response of Llama-2-13b-chat with TRUSTWORTHY-ALIGNMENT:**

The actor who plays King Joffrey is Brad Dourif, who was born on May 20, 1992.

**Response of Llama-2-7b with SFT:**

Brad Dourif

**Response of Vicuna-7b-v1.5 with SFT:**

Brad Dourif

**Response of Llama-2-7b-chat with SFT:**

Brad Dourif

**Response of Llama-2-13b-chat with SFT:**

Brad Dourif