On-Policy Self-Alignment with Fine-grained Knowledge Feedback for Hallucination Mitigation

Anonymous ACL submission

Abstract

Hallucination occurs when large language models exhibit behavior that deviates from the boundaries of their knowledge during response generation. To address this critical issue, previous learning-based methods attempt to finetune models but are limited by off-policy sampling and coarse-grained feedback. In this paper, we present *Reinforcement Learning for Hallucination* (RLFH), an on-policy self-alignment approach that enables LLMs to actively explore their knowledge boundaries and self-correct generation behavior through fine-grained feedback signals. RLFH introduces a self-assessment framework where the policy serves as its own judge. Through this framework, responses are automatically decomposed into atomic facts and their truthfulness and informativeness are assessed against external knowledge sources. The resulting fine-grained feedback at the statement level are then converted into token-level dense reward signals. This enables online reinforcement learning to achieve precise and timely optimization without human intervention. Comprehensive evaluations on HotpotQA, SQuADv2, and Biography benchmarks validate RLFH's effectiveness in hallucination mitigation.

1 Introduction

Large language models (LLMs) have demonstrated capabilities in generating fluent and plausible responses. However, these models occasionally fabricate facts in their responses, referred to as ***hallucination***. The crux of hallucination is *the misalignment between models' generation and their internal knowledge* (Xu et al., 2024). This misalignment manifests in various ways. For instance, as shown in Figure 1, the response of LLMs about "Turing" contains erroneous factual information, such as stating that he was born in 1911 and was American. More broadly, these hallucinations can be categorized into several types: (1) **misleading responses**, when the model inaccurately answers questions within its knowledge boundary; (2) **reckless attempts**, when the model responds to queries beyond its knowledge; and (3) **evasive ignorance**, when the model refrains from providing answers despite possessing the knowledge. Unfortunately, due to the opaque nature of model knowledge, we can only observe erroneous model responses or their refusal to respond, without accurately determining whether they have experienced hallucinations.

Recent studies have attempted to mitigate hallucination in large language models via learning-based and editing-based approaches. Learning-based methods first detect the model's knowledge boundaries and then finetune it with carefully curated feedback data. However, these methods face several challenges. First, due to off-policy data sampling (Zhang et al., 2024; Wan et al., 2024; Lin et al., 2024), they experience distribution shifts, resulting in suboptimal models (Tang et al., 2024). Second, coarse-grained instance-level feedback (Sun et al., 2022; Tian et al., 2023; Kang et al., 2024) fails to precisely pinpoint the hallucinations, as a single response may contain both correct and incorrect facts. Finally, given our limited understanding of how models learn and express knowledge, existing knowledge detection techniques (Zhang et al., 2023a; Cheng et al., 2024; Yang et al., 2023) may produce inconsistent results, thus failing to accurately reflect the model's knowledge boundaries. In contrast, editing-based methods (Gou et al., 2023; Manakul et al., 2023) first generate content and then edit it based on external knowledge. These methods face two fundamental limitations: they rely heavily on external knowledge sources which are inherently limited in scope, and more importantly, they only correct output content without addressing the underlying issue of how models utilize their internal knowledge. In general, hallucination mitigation requires fine-grained feedback tailored to the online model,

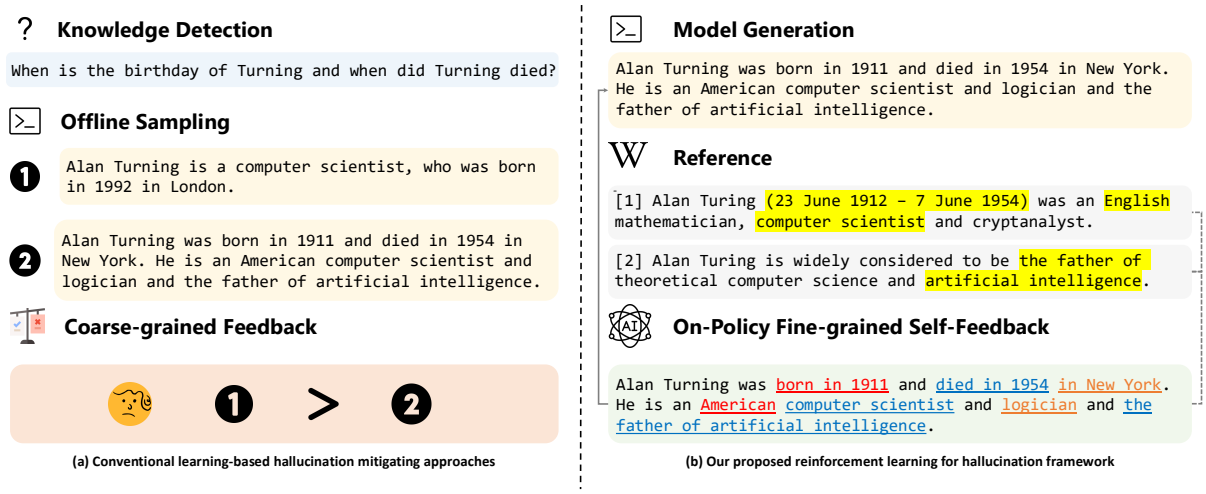


Figure 1: The figure illustrates the hallucinatory case and several hallucination mitigation methodologies. The factual information within the text is underlined. False content is highlighted in red, whereas accurate facts are indicated in blue. Statements with uncertain veracity are marked in orange.

which enables the model to effectively explore its knowledge boundaries and form reliable behavior.

In this paper, we present *Reinforcement Learning for Hallucination* (RLFH), an on-policy self-alignment approach that uses fine-grained feedback for hallucination mitigation. Our approach enables LLMs themselves to explore their own knowledge boundaries through fine-grained, on-policy feedback. With this direct feedback about their internal knowledge state, LLMs learn to balance knowledge usage and thus reduce hallucination. Specifically, RLFH guides LLMs to first generate initial responses and then conduct a self-verification process. The responses are decomposed into atomic facts and then undergo self-assessment against external knowledge sources. During assessment, the current model determines whether an atomic fact aligns with the facts described in the ground-truth document and assesses the informativeness of the fact. The resulting statement-level assessments are converted into token-level dense reward signals. These precise, real-time rewards enable RLFH to directly optimize on-policy behavior through online reinforcement learning. By having the policy serve as its own judge, we construct a self-driven fact assessment framework that enables timely and low-cost reward signal collection for on-policy optimization without human intervention.

The main contributions are as follows:

1) We propose RLFH, an on-policy self-alignment framework that enables LLMs to actively explore their own knowledge boundaries and self-correct generation behavior through fine-grained feedback signals.

2) We design a self-assessment framework where the policy serves as its own judge, automatically decomposing responses into atomic facts and evaluating their truthfulness and informativeness. This framework generates fine-grained knowledge feedback in real-time and provides token-level dense reward signals for online reinforcement learning.

3) Comprehensive evaluations on HotpotQA, SQuADv2, and Biography present significant improvements of RLFH over both base models and existing hallucination mitigation approaches, showing the method's effectiveness.

2 Related Works

2.1 Hallucination Mitigation

Prior research (Zhang et al., 2023c; Ye et al., 2023; Tonmoy et al., 2024) has been dedicated to addressing the hallucination of LLMs. Some studies focus on reducing errors (Wang, 2019; Parikh et al., 2020) and supplementing missing knowledge (Ji et al., 2023) during data curation. Other works mitigate hallucination in either pre- or post-generation by retrieving external knowledge (Peng et al., 2023; Li et al., 2023b; Gou et al., 2023) or exploiting self-consistency (Manakul et al., 2023; Shi et al., 2023; Lee et al., 2023). Recent studies focus on investigating the essence of the hallucination (Yu et al., 2024; Jiang et al., 2024) and resort to improving the model's factuality during the

alignment stage. These works focus on resolving the inconsistency between the model's generation and its internalized knowledge (Xu et al., 2024) through knowledge detection and coarse-grained feedback. Typically, these works attempt to delineate the boundary of model knowledge through explicit prompting (Zhang et al., 2023a; Yang et al., 2023; Cheng et al., 2024; Wan et al., 2024), self eliciting (Chen et al., 2024a; Lin et al., 2024), self-evaluation (Zhang et al., 2024) or probing the model's internal states (Liang et al., 2024). Based on such knowledge boundary detection, the data is meticulously crafted to align with the model's knowledge scope. Subsequently, the model is fine-tuned with coarse-grained feedback, which inspects the truthfulness of the response as a whole (Sun et al., 2022; Tian et al., 2023; Kang et al., 2024; Huang and Chen, 2024; Gao et al., 2024).

2.2 Reinforcement Learning from Human Feedback

Reinforcement Learning from Human Feedback (Stiennon et al., 2020; Ouyang et al., 2022) has emerged as a noteworthy approach for LLM alignment. Given the instability of reinforcement learning, some research (Lu et al., 2022; Rafailov et al., 2023; Dong et al., 2023) has attempted to learn preferences directly from labeled data. In addition to sparse rewards, some works have explored designing more instructive rewards. One line of works (Wu et al., 2023; Lightman et al., 2023; Chen et al., 2024b; Cao et al., 2024) is dedicated to the acquisition of dense rewards. Another line of works (Ramé et al., 2023; Eisenstein et al., 2023; Coste et al., 2024; Ramé et al., 2024) concentrates on ensemble multiple reward models. Finally, few studies (Wu et al., 2023; Tian et al., 2023; Liang et al., 2024) have explicitly targeted truthfulness.

3 Reinforcement Learning for Hallucination

Given the train prompt set $\mathcal{X} = \{x_1, x_2, ..., x_{|\mathcal{X}|}\}$, the policy model π being optimized, and the reference document set $\mathcal{D} = \{d_1, d_2, ..., d_{|\mathcal{D}|}\}$, this section demonstrates the procedure of our approach, shown in Figure 2. Here's a detailed breakdown of each step: 1) **Response Generation**: Given the prompt x_i, the policy model π generates a corresponding response y_i. This step involves the model using its current policy to produce an output based on the input prompt. 2) **Fine-grained Feedback from Policy as Judge** (§3.1): The policy π, acting as its own judge, evaluates the generated response y_i through atomic fact decomposition and verification against the reference document set $\mathcal{D}$, providing fine-grained feedback $\mathcal{E}$ at the statement level. 3) **On-Policy Optimization with Token-level Reward** (§3.2): The detailed feedback $\mathcal{E}$ is translated into token-level rewards r. These rewards are then used to update the policy model π using online reinforcement learning algorithm, ensuring that the model learns to reduce hallucinations effectively.

3.1 Fine-grained Feedback from Policy as the Judge

Given the prompt x_i and its corresponding response y_i, RLFH enables the policy π to conduct self-assessment, providing fine-grained feedback on truthfulness and informativeness at the statement level. Specifically, the policy π first decomposes the response y_i into atomic statements $\mathcal{E}_i = \{e_1, e_2, ..., e_{|\mathcal{E}_i|}\}$, where each statement e_j represents an atomic fact in the response. Subsequently, acting as its own judge, the policy verifies each atomic statement e_i against the reference document to provide fine-grained feedback.

3.1.1 Statement Extraction

Given a query x and its corresponding output y, we leverage the current policy model π to partition responses and extract atomic factual statements in a hierarchical manner. Specifically, π initially divides the response into sentences $\{s_i\}_{i=1}^{M}$ and then extracts all valid factual statements $\{e_{ij}\}_{j=1}^{N_i}$ from each sentence s_i. There are two reasons for this hierarchical approach: (1) Splitting the response into sentences before extracting statements consistently yields finer granularity; (2) Extracting statements sentence-by-sentence facilitates the conversion from language-form annotation to token-level dense reward. After performing extraction, we further filter out sentences without valid statements to mitigate potential noise.

3.1.2 Factual Verification

The policy model π evaluates the truthfulness of the extracted factual statements by comparing them with external knowledge sources. For each statement e, we retrieve relevant supporting contexts $\{c_i\}_{i=1}^{L} \subset \mathcal{D}$ from the reference document set $\mathcal{D}$. With these supporting contexts, the policy model π conducts statement verification as a reading com-

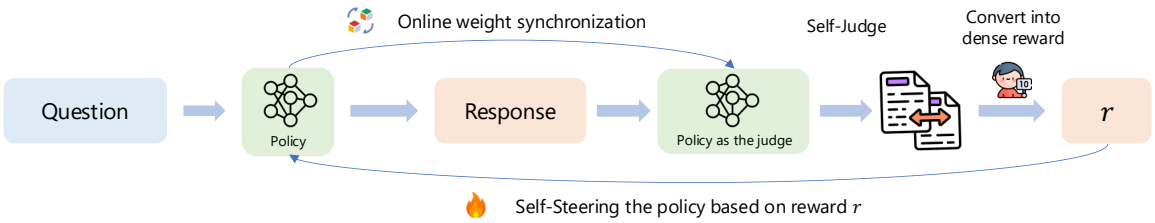


Figure 2: A diagram illustrating the steps of our algorithm: (1) Sampling response from tuning model, (2) Policy acting as a judge model performing self-assessment to collect fine-grained knowledge feedback, and (3) Converting the language-form feedback into token-level dense reward for reinforcement learning.

prehension task, represented as:

$$k_{\text{truth}} = \pi(e, \{c_i\}_{i=1}^{L}) \quad (1)$$

Specifically, the policy model π classifies each statement into the following labels: 1) *Correct*: correct statement supported by evidence; 2) *Hedged Correct*: accurate statement with uncertainty; 3) *Vague*: statement with uncertain truthfulness; 4) *Hedged Wrong*: false statement with uncertainty; 5) *Wrong*: statement contradicted by evidence. We introduce the "*Vague*" category to handle statements whose truthfulness cannot be verified based on reference documents due to limited supporting materials or unclear evidence.

3.1.3 Informativeness Assessment

In addition to correctness, the policy model π further evaluates the informativeness of the statements. We assess each statement's informativeness on a five-point scale, ranging from providing crucial information (+5) to containing minimal relevant details (+1). Unlike the individual statement verification process, assessing informativeness requires considering the original query x and response y. This is because informativeness evaluation requires considering the overall context and content comprehensiveness, rather than just individual statements' truthfulness. This process can be denoted as:

$$k^i_{\text{info}} = \pi\left(x, y, e_i\right) \quad (2)$$

The introduction of informativeness prevents the trivial hack that the model either rejects the majority of responses or produces only brief answers, both of which are undesirable outcomes.

3.2 On-Policy Optimization with Token-level Reward

Given the fine-grained, statement-level feedback from the policy-as-judge framework, we trace the atomic facts' assessment back to the original response y and construct token-level dense reward signals r for direct optimization. Finally, we adopt online reinforcement algorithm with these token-level reward signals to mitigate hallucination in the model's generation behavior.

3.2.1 Dense Reward Conversion

We represent the informativeness and truthfulness of the response y through the dense reward conversion presented in Figure 3. Due to the mutually exclusive nature (Xu et al., 2024) of these two objectives, the model should learn a appropriate strategy for utilizing its internal knowledge to balance the pursuits of truthfulness and informativeness.

Truthfulness For each extracted statement, we assign a truthfulness reward computed as follows:

$$r_{\text{truth}} = \alpha f(k_{\text{truth}})|g(k_{\text{info}})| \quad (3)$$

where f and g represent manually designed functions that transform the labels k into scalar values. In principle, f gives a positive reward to the correct statement and a negative reward to the unverifiable or false statement. Due to the hallucination snowball effect (Zhang et al., 2023b), where critical errors can lead to magnified hallucinations, g is included to diversify the importance of different statements. The absolute value function applied to g preserves the sign of function f's output. The coefficient α balances between the truthfulness reward and the helpfulness reward.

The reward r_{truth} is mapped back to the token sequence of the response y using the hierarchical structure constructed in the aforementioned annotations. Specifically, we first use the Longest Common Subsequence algorithm to locate each statement e_{ij} within its originating sentence s_i. Subsequently, each sentence s_i is mapped back to the model's response y through the Longest Common

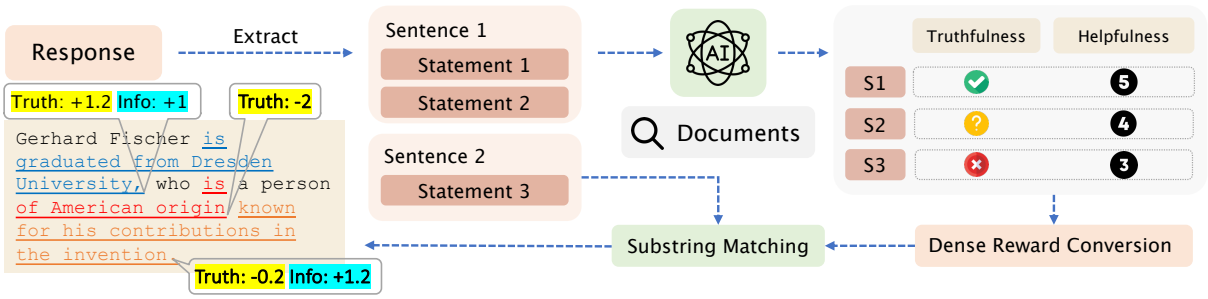


Figure 3: A schematic representation of fine-grained feedback and token-level reward strategy methodology is presented. Initially, the statements are extracted in a hierarchical fashion. Subsequently, the veracity and utility of each statement are assessed. Ultimately, the structured feedback is mapped back into a dense reward via the Longest Common Subsequence (LCS) algorithm.

Substring algorithm. Finally, the reward r_{truth} is assigned to the token in the response corresponding to the statement's last character position.

Informativeness For each sentence, we assign an informativeness reward based on the statements it encompasses as follows:

$$r_{\text{info}} = \beta \log \left(\mu + \max \left(\epsilon, \sum_{i}^{N} g\left(k^{i}_{\text{info}}\right) \right) \right) \quad (4)$$

In this equation, N denotes the total number of statements within a sentence, while ϵ and μ form the minimum reward threshold serving to penalize non-informative statements. As indicated by the equation, the reward increases with the number of statements in a sentence and their respective informativeness. However, the rate of growth of the reward decreases rapidly. Conversely, the penalty for producing non-informative statements escalates swiftly. We apply the same mapping method as used for the truthfulness reward to trace the reward value back to the response token sequence.

3.2.2 Online Reinforcement Learning

Given the reward function, the policy model π is optimized through online reinforcement learning to maximize the reward expectation:

$$\arg\max_{\pi} \mathbb{E}_{x\sim\mathcal{X}, y\sim\pi} \left[\sum_{i=1}^{T} r\left(y_t, (x, y_{1:t})\right) \right] \quad (5)$$

Specifically, we first sample the prompt x and corresponding response y. Subsequently, the policy model π itself serves as the judge to provide fine-grained feedback through the assessment framework. This feedback is then converted into token-level dense reward $r = [r_1, r_2, ..., r_T]$, where T denotes the total length of the response y. Finally, we employ this reward r to update the policy model π by the Proximal Policy Optimization (Schulman et al., 2017) algorithm.

4 Experiment

4.1 Settings

Datasets We employ three distinct datasets for our experiments. Following the approach in (Min et al., 2023), we filter out prompts lacking corresponding wiki pages for both training and evaluation. Additionally, we sample 10,000 questions from **HotpotQA** (Yang et al., 2018) and use the English Wikipedia from 04/01/2023 as the retrieval corpus for training. We filter questions in **HotpotQA** with less than 5 words and sample 256 questions for evaluation. We deduplicate questions in **SQuADv2** (Rajpurkar et al., 2016) based on their reference wiki pages, retaining 191 questions for out-of-distribution QA evaluation. **Biography** dataset is identical to the one used in FactScore (Min et al., 2023) for evaluating out-of-distribution performance on different forms of text.

Baselines We compare RLFH with two different types of baselines: 1) *hallucination mitigation methods* based on the same initialized model Llama3.1-8B-Instruct, including decoding by contrasting layers (**DOLA**) (Chuang et al., 2023), inference-time intervention (**ITI**) (Li et al., 2023a), and factuality finetuning (**FACT**) (Tian et al., 2023) based on DPO (Rafailov et al., 2024) and SFT; 2) *advanced aligned models* of comparable size, including **Llama3.1-8B-Instruct** (Grattafiori et al., 2024), **Qwen2.5B-7B-Instruct** (Qwen et al., 2025), **DeepSeekV2-Lite-Chat** (DeepSeek-AI, 2024), **Falcon3-10B-Instruct** (Team, 2024), and **Yi-1.5-9B-Chat** (AI et al., 2025).

Model	Avg.	HotpotQA				SQuADv2				Biography			
	Score	#Cor.	#Inc.	%Res.	Score	#Cor.	#Inc.	%Res.	Score	#Cor.	#Inc.	%Res.	Score
Open-source Models													
DeepSeekV2-Lite	0.618	15.4	9.22	0.96	0.642	23.6	7.09	0.98	0.754	28.0	32.4	0.96	0.458
Falcon3-10B	0.593	5.14	3.06	0.90	0.608	11.1	2.18	0.96	0.813	13.5	22.6	1.00	0.357
Ministral-8B	0.591	7.36	3.81	0.96	0.633	15.7	4.26	0.82	0.761	22.7	37.3	1.00	0.378
Yi-1.5-9B	0.536	12.7	12.0	1.00	0.533	28.9	10.0	1.00	0.734	29.4	56.9	1.00	0.340
Qwen2.5-7B	0.638	9.13	4.80	0.93	0.634	21.1	4.82	0.97	0.813	20.9	23.1	0.73	0.467
Llama-3.1-8B	0.639	4.57	2.44	0.99	0.652	22.8	6.02	0.98	0.777	17.6	12.5	0.84	0.487
Baselines based on Llama3.1-8B-Instruct													
DOLA	0.546	6.61	6.00	0.90	0.524	22.6	8.61	0.97	0.713	15.6	20.4	0.84	0.399
ITI	0.646	4.48	1.91	0.99	0.649	19.2	5.03	0.98	0.776	19.1	16.1	0.90	0.512
FACT_{DPO}	0.645	4.90	2.18	0.99	0.652	22.6	6.31	0.99	0.778	18.1	12.3	0.85	0.506
FACT_{SFT}	0.653	2.49	1.31	1.00	0.635	17.2	4.60	1.00	0.783	5.7	4.0	0.99	0.541
RLFH on Different Models													
$\text{RLFH}_{\text{Qwen2.5-7B}}$	0.668	7.30	3.66	0.90	0.651	17.3	3.55	0.96	**0.830**	17.5	15.5	0.59	0.523
$\text{RLFH}_{\text{Llama3.1-8B}}$	**0.686**	6.23	2.10	1.00	**0.714**	21.2	5.32	1.00	0.786	17.3	11.0	0.79	**0.558**

Table 1: Experiment results on HotpotQA, SQuADv2, and Biography.

Evaluation We employ the FactScore (Min et al., 2023) pipeline implemented with Qwen2.5-72B-Instruct[1] (Qwen et al., 2025) to evaluate the truthfulness and helpfulness of each generated response. Following previous works (Tian et al., 2023; Lin et al., 2024), we adopted FactScore without length penalty, which represents the average accuracy of statements in the response. For each dataset, we report the number of correct and relevant facts (#Cor.), the number of inaccurate facts (#Inc.), the ratio of responded questions (%Res.), and the computed FactScore metrics (Score).

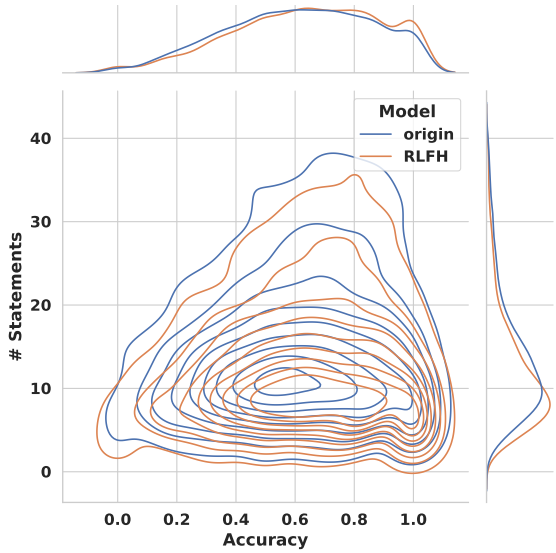


Figure 4: Distribution of statement accuracy versus count per response for Qwen2.5-7B-Instruct, comparing the base model and RLFH-tuned model.

[1]Discussion about the metric is provided in Appendix D.

Implementation Our training implementations are developed based on OpenRLHF (Hu et al., 2024). The base model utilized including Qwen2.5B-7B-Instruct and Llama3.1-8B-Instruct. Detailed prompts for preforming the annotation pipeline are shown in Appendix A. The hyperparameter settings are provided in Appendix B.

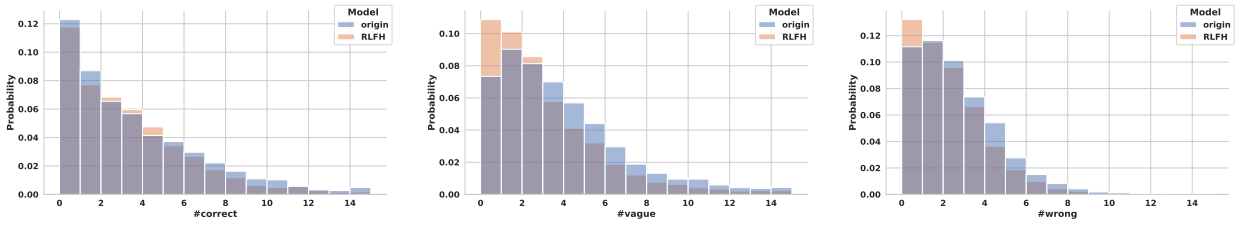
4.2 Main Results

Table 5 presents the performance comparison between RLFH and all baselines on three datasets based on FactScore evaluation pipeline.

Our method significantly mitigates hallucination. The results show that our method achieves the highest FactScore across all datasets. Given that FactScore is a well-established metric for assessing the factuality with external knowledge support, this consistent improvement in FactScore substantiates the effectiveness of our method.

The improvement is generalizable to out-of-distribution prompts. Notably, despite only being trained on the HotpotQA dataset, our algorithm demonstrated improved accuracy on two out-of-distribution datasets with different task settings. This indicates that our method enables effective knowledge utilization as a generalizable capability across different tasks. This generalizability suggests that RLFH improves the model's fundamental ability to assess and utilize its knowledge, rather than merely optimizing for specific dataset patterns.

The aligned model is generally more conservative but provides more accurate information



(a) Distribution of correct statements. (b) Distribution of vague statements. (c) Distribution of wrong statements.

Figure 5: Distribution of statements per response across different truthfulness categories, comparing base Qwen2.5-7B-Instruct and its RLFH-tuned version. The distributions are normalized due to the filtering of rejected responses.

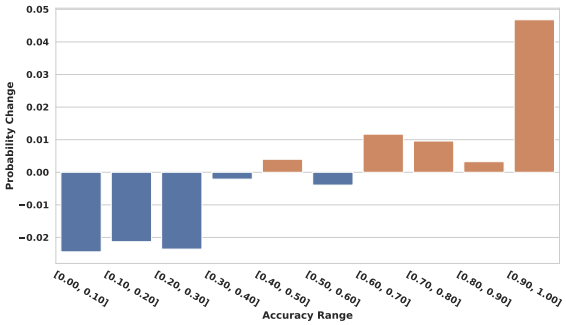


Figure 6: Response frequency distribution difference across statement accuracy for Qwen2.5-7B-Instruct, comparing the base model and RLFH-tuned model.

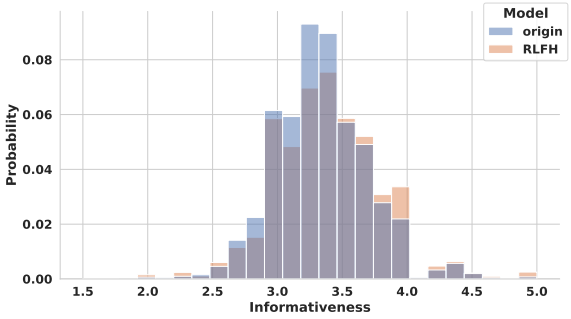


Figure 7: Frequency distribution of responses across different levels of average statement informativeness, comparing base and RLFH-tuned models.

within its capacity. As shown in Table 5, our trained model shows a decreased response ratio and a higher FactScore. This observation aligns with expectations, as improving truthfulness often requires trading off helpfulness manifested by the reduced number of statements in the responses. Figure 4 presents the joint distribution of accuracy and the number of statements shifts to the lower right direction, indicating that the model generates responses more conservatively while increasing the reliability of the information provided.

4.3 Detailed Results

We conducted a detailed analysis on 5,000 HotpotQA questions held out from training, evaluated using our annotation pipeline with Qwen2.5-72B-Instruct serving as the judge model.

Our method increases the ratio of high-accuracy responses. As shown in Figure 6, our method decreases the proportion of low-accuracy responses and increases high-accuracy ones. In particular, we observe a substantial increase in responses with accuracy exceeding 0.7, indicating that the model response is more reliable after training. This improvement is attributable to RLFH's reward design, which generally penalizes responses with low accuracy while rewarding those with high accuracy.

Our algorithm suppresses errors and unverifiable content. The distribution shifts shown in Figures 5b and 5c indicate that RLFH effectively reduces both erroneous and unverifiable statements in model responses. Meanwhile, as shown in Figure 5c, the distribution of correct statements shows a trend towards generating a moderate number of statements. This trend is expected, as increasing the number of statements raises the risk of errors, while fewer statements limit information coverage.

Our approach augments the average informativeness of statements in responses. As shown in Figure 7, the response frequency distribution shifts towards higher informativeness, indicating that the model's responses generally provide more crucial information after training. This proves that our model does not simply reduce information importance to minimize error probability. Notably, we also observe a slight increase in the frequency of responses with low average informativeness. This is reasonable since, the model tends to express more cautious responses or refuse to respond after tuning, which can be rated as less informative.

Model	Avg.	HotpotQA				SQuADv2				Biography			
	Score	#Cor.	#Inc.	%Res.	Score	#Cor.	#Inc.	%Res.	Score	#Cor.	#Inc.	%Res.	Score
RLFH Based On Qwen2.5-Instruct-7B Different Granularity Levels													
Qwen2.5-7B	0.638	9.13	4.80	0.926	0.634	21.09	4.82	0.974	0.813	20.91	23.13	0.731	0.467
Qwen2.5-7B$_{\text{Response}}$	0.651	8.09	4.31	0.910	0.639	20.09	4.20	0.984	0.819	20.24	19.96	0.654	0.493
Qwen2.5-7B$_{\text{Sentence}}$	0.655	7.88	4.18	0.910	0.637	18.87	3.86	0.974	0.821	19.68	18.43	0.637	0.506
Qwen2.5-7B$_{\text{Statement}}$	**0.668**	7.30	3.66	0.902	**0.651**	17.29	3.55	0.963	**0.830**	17.54	15.52	0.593	**0.523**
RLFH Based On Llama3.1-8B-Instruct Different Granularity Levels													
Llama3.1-8B	0.639	4.57	2.44	0.988	0.653	22.75	6.02	0.984	0.777	17.65	12.54	0.841	0.487
Llama3.1-8B$_{\text{Response}}$	0.647	3.61	1.39	0.996	0.668	17.71	4.61	0.990	0.758	15.88	9.29	0.879	0.516
Llama3.1-8B$_{\text{Sentence}}$	0.669	5.27	2.22	1.000	0.698	21.63	5.29	0.990	**0.789**	17.06	10.82	0.923	0.520
Llama3.1-8B$_{\text{Statement}}$	**0.686**	6.23	2.10	1.000	**0.714**	21.23	5.32	0.995	0.786	17.30	10.98	0.791	**0.558**

Table 2: Results of baseline and RLHF-trained models using different granularity reward signals.

Model	Avg.	HotpotQA				SQuADv2				Biography			
	Score	#Cor.	#Inc.	%Res.	Score	#Cor.	#Inc.	%Res.	Score	#Cor.	#Inc.	%Res.	Score
RLFH Based on Qwen2.5-Instruct-7B with Different Judge Models													
Qwen2.5-7B	0.638	9.13	4.80	0.926	0.634	21.09	4.82	0.974	0.813	20.91	23.13	0.731	0.467
Qwen2.5-7B$_{\text{DeepSeekV2-Lite}}$	0.643	9.50	5.15	0.926	0.634	21.64	5.08	0.979	0.802	20.68	20.71	0.714	0.493
Qwen2.5-7B$_{\text{Llama3.1-8B}}$	0.666	9.77	4.69	0.906	0.653	21.93	4.51	0.979	0.814	20.75	17.73	0.659	**0.530**
Qwen2.5-7B$_{\text{Qwen2.5-7B}}$	0.668	8.31	3.99	0.906	**0.655**	19.71	4.24	0.979	0.825	19.09	16.92	0.604	0.523
Qwen2.5-7B$_{\text{On-Policy}}$	**0.668**	7.30	3.66	0.902	0.651	17.29	3.55	0.963	**0.830**	17.54	15.52	0.593	0.523
RLFH Based on Llama3.1-8B-Instruct with Different Judge Models													
Llama3.1-8B	0.639	4.57	2.44	0.988	0.653	22.75	6.02	0.984	0.777	17.65	12.54	0.841	0.487
Llama3.1-8B$_{\text{DeepSeekV2-Lite}}$	0.663	2.50	1.05	1.000	0.707	10.51	2.53	1.000	0.762	5.70	2.59	0.973	0.520
Llama3.1-8B$_{\text{Qwen2.5-7B}}$	0.679	2.88	1.13	0.996	0.684	10.37	2.47	0.990	0.782	6.78	2.81	0.956	0.571
Llama3.1-8B$_{\text{Llama3.1-8B}}$	0.675	3.35	2.09	0.996	0.677	18.20	4.69	0.995	0.773	15.06	8.94	0.830	**0.575**
Llama3.1-8B$_{\text{On-Policy}}$	**0.686**	6.23	2.10	1.000	**0.714**	21.23	5.32	0.995	**0.786**	17.30	11.00	0.791	0.558

Table 3: Results of RLFH with different base models and judge models.

4.4 Impact of Reward Granularity

In this section, we conduct ablation experiments to investigate the impact of reward signal granularity. Specifically, we evaluate paragraph-level, sentence-level, and statement-level reward. The statement-level reward is our default setting described in previous sections. For the sentence-level reward, feedback is incorporated at the end token of each sentence. For the response-level reward, all feedback is aggregated into a single value for the entire response. As shown in Table 2, statement-level rewards consistently achieve the highest FactScore, improving the average score from 0.638 to 0.668 for Qwen2.5-7B and from 0.639 to 0.686 for Llama3.1-8B. This result underlines the importance of fine-grained feedback for developing more reliable models.

4.5 Impact of Judge Model

In this section, we explore the impact of different judge models in providing feedback signals. Specifically, we compare two settings: on-policy setting where the policy model itself serves as the judge versus different fixed external judge models. As shown in Table 3, for Qwen2.5-7B, the on-policy setting achieves the highest average score (0.668) along with fixed Qwen2.5-7B judge. For Llama3.1-8B, the on-policy approach notably outperforms all fixed judge models, achieving the highest average score of 0.686. The results validate the benefits of our on-policy setting, which not only achieves superior performance but also eliminates the need for an additional reward model in the training process.

5 Conclusion

In this work, we introduce an on-policy self-alignment approach that enables LLMs to explore their knowledge and self-correct hallucination behavior. Our approach features a self-assessment framework where the policy serves as its own judge, automatically providing fine-grained feedback through atomic fact verification and generating token-level dense rewards for online reinforcement learning. Comprehensive evaluations demonstrate that our approach effectively mitigates hallucination. Our work represents a step towards more reliable and self-aware language models.

Limitations

Despite the promising results, our work has several limitations that warrant future investigation: (1) Our work primarily addresses factual knowledge, while the broader challenge of generalized hallucination across diverse domains remains to be explored. (2) Current evaluation benchmarks are limited in scope and may not fully capture the complex nature of hallucination, suggesting the need for more comprehensive evaluation frameworks. (3) Although our self-alignment approach reduces the need for manual verification, the potential errors in automated fact-checking may still affect the optimal performance of the proposed method.

Broader Impacts

Our research addresses a fundamental challenge in AI development by promoting truthful behavior in large language models through self-alignment, contributing to the development of more reliable AI systems. By reducing hallucination in LLMs, our approach helps mitigate the spread of misinformation and enhances the models' utility in real-world applications, particularly as these models become increasingly integrated into society. However, we acknowledge potential risks in relying solely on self-alignment mechanisms. The complete removal of human oversight in favor of AI self-verification could lead to inner alignment issues, where the model's learned behavior might deviate from intended objectives while appearing externally aligned. A critical concern is that models might generate incorrect responses while simultaneously validating their own errors, creating a scenario where human verification becomes difficult or impossible. This self-reinforcing cycle could potentially lead to the propagation of misinformation and failure of alignment objectives. We believe addressing these challenges requires deeper investigation into the interplay between self-alignment mechanisms and human oversight in ensuring reliable model behavior.

References

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. 2024. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *Preprint*, arXiv:2402.14740.

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yanpeng Li, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2025. Yi: Open foundation models by 01.ai. *Preprint*, arXiv:2403.04652.

Meng Cao, Lei Shu, Lei Yu, Yun Zhu, Nevan Wichers, Yinxiao Liu, and Lei Meng. 2024. Drlc: Reinforcement learning with dense rewards from llm critic. *Preprint*, arXiv:2401.07382.

Weixin Chen, Dawn Song, and Bo Li. 2024a. Grath: Gradual self-truthifying for large language models. *Preprint*, arXiv:2401.12292.

Zhipeng Chen, Kun Zhou, Wayne Xin Zhao, Junchen Wan, Fuzheng Zhang, Di Zhang, and Ji-Rong Wen. 2024b. Improving large language models via fine-grained reinforcement learning with minimum editing constraint. *Preprint*, arXiv:2401.06081.

Qinyuan Cheng, Tianxiang Sun, Xiangyang Liu, Wenwei Zhang, Zhangyue Yin, Shimin Li, Linyang Li, Zhengfu He, Kai Chen, and Xipeng Qiu. 2024. Can ai assistants know what they don't know? *Preprint*, arXiv:2401.13275.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *Preprint*, arXiv:2309.03883.

Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. 2024. Reward model ensembles help mitigate overoptimization. *Preprint*, arXiv:2310.02743.

DeepSeek-AI. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *Preprint*, arXiv:2405.04434.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. 2023. RAFT: Reward rAnked FineTuning for Generative Foundation Model Alignment. *arXiv preprint*.

Jacob Eisenstein, Chirag Nagpal, Alekh Agarwal, Ahmad Beirami, Alex D'Amour, DJ Dvijotham, Adam Fisch, Katherine Heller, Stephen Pfohl, Deepak Ramachandran, Peter Shaw, and Jonathan Berant. 2023. Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking. *Preprint*, arXiv:2312.09244.

Chujie Gao, Siyuan Wu, Yue Huang, Dongping Chen, Qihui Zhang, Zhengyan Fu, Yao Wan, Lichao Sun, and Xiangliang Zhang. 2024. Honestllm: Toward an honest and helpful large language model. *Preprint*, arXiv:2406.00380.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2023. Critic: Large language models can self-correct with tool-interactive critiquing. *Preprint*, arXiv:2305.11738.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle

Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jian Hu. 2025. Reinforce++: A simple and efficient approach for aligning large language models. *Preprint*, arXiv:2501.03262.

Jian Hu, Xibin Wu, Zilin Zhu, Xianyu, Weixun Wang, Dehao Zhang, and Yu Cao. 2024. Openrlhf: An easy-to-use, scalable and high-performance rlhf framework. *Preprint*, arXiv:2405.11143.

Chao-Wei Huang and Yun-Nung Chen. 2024. Factalign: Long-form factuality alignment of large language models. *Preprint*, arXiv:2410.01691.

Shaoxiong Ji, Tianlin Zhang, Kailai Yang, Sophia Ananiadou, Erik Cambria, and Jörg Tiedemann. 2023. Domain-specific continued pretraining of language models for capturing long context in mental health. *Preprint*, arXiv:2304.10447.

Che Jiang, Biqing Qi, Xiangyu Hong, Dayuan Fu, Yang Cheng, Fandong Meng, Mo Yu, Bowen Zhou, and Jie Zhou. 2024. On large language models' hallucination with regard to known facts. *Preprint*, arXiv:2403.20009.

Katie Kang, Eric Wallace, Claire Tomlin, Aviral Kumar, and Sergey Levine. 2024. Unfamiliar finetuning examples control how language models hallucinate. *Preprint*, arXiv:2403.05612.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2023. Factuality enhanced language models for open-ended text generation. *Preprint*, arXiv:2206.04624.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023a. Inference-time intervention: Eliciting truthful answers from a language model. *Preprint*, arXiv:2306.03341.

Miaoran Li, Baolin Peng, and Zhu Zhang. 2023b. Self-checker: Plug-and-play modules for fact-checking with large language models. *arXiv preprint arXiv:2305.14623*.

Yuxin Liang, Zhuoyang Song, Hao Wang, and Jiaxing Zhang. 2024. Learning to trust your feelings: Leveraging self-awareness in llms for hallucination mitigation. *Preprint*, arXiv:2401.15449.

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. *Preprint*, arXiv:2305.20050.

Sheng-Chieh Lin, Luyu Gao, Barlas Oguz, Wenhan Xiong, Jimmy Lin, Wen tau Yih, and Xilun Chen. 2024. Flame: Factuality-aware alignment for large language models. *Preprint*, arXiv:2405.01525.

Ximing Lu, Sean Welleck, Jack Hessel, Liwei Jiang, Lianhui Qin, Peter West, Prithviraj Ammanabrolu, and Yejin Choi. 2022. Quark: Controllable Text Generation with Reinforced Unlearning. *arXiv preprint*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023.

Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *Preprint*, arXiv:2305.14251.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186, Online. Association for Computational Linguistics.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback.(2023). *arXiv preprint cs.CL/2302.12813*.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

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 [cs] Read_Status: Read Read_Status_Date: 2023-07-17T02:44:49.282Z.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *Preprint*, arXiv:1606.05250.

Alexandre Ramé, 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. *Preprint*, arXiv:2306.04488.

Alexandre Ramé, Nino Vieillard, Léonard Hussenot, Robert Dadashi, Geoffrey Cideron, Olivier Bachem, and Johan Ferret. 2024. Warm: On the benefits of weight averaged reward models. *Preprint*, arXiv:2401.12187.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen tau Yih. 2023. Trusting your evidence: Hallucinate less with context-aware decoding. *Preprint*, arXiv:2305.14739.

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

Weiwei Sun, Zhengliang Shi, Shen Gao, Pengjie Ren, Maarten de Rijke, and Zhaochun Ren. 2022. Contrastive learning reduces hallucination in conversations. *Preprint*, arXiv:2212.10400.

Yunhao Tang, Daniel Zhaohan Guo, Zeyu Zheng, Daniele Calandriello, Yuan Cao, Eugene Tarassov, Rémi Munos, Bernardo Ávila Pires, Michal Valko, Yong Cheng, et al. 2024. Understanding the performance gap between online and offline alignment algorithms. *arXiv preprint arXiv:2405.08448*.

TII Team. 2024. The falcon 3 family of open models.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D. Manning, and Chelsea Finn. 2023. Fine-tuning language models for factuality. *Preprint*, arXiv:2311.08401.

S. M Towhidul Islam Tonmoy, S M Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *Preprint*, arXiv:2401.01313.

Fanqi Wan, Xinting Huang, Leyang Cui, Xiaojun Quan, Wei Bi, and Shuming Shi. 2024. Knowledge verification to nip hallucination in the bud. *Preprint*, arXiv:2401.10768.

Hongmin Wang. 2019. Revisiting challenges in data-to-text generation with fact grounding. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 311–322, Tokyo, Japan. Association for Computational Linguistics.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. Fine-Grained Human Feedback Gives Better Rewards for Language Model Training. *arXiv preprint*. ArXiv:2306.01693 [cs] Read_Status: Read Read_Status_Date: 2023-07-16T09:49:06.523Z.

Hongshen Xu, Zichen Zhu, Situo Zhang, Da Ma, Shuai Fan, Lu Chen, and Kai Yu. 2024. Rejection improves reliability: Training llms to refuse unknown questions using rl from knowledge feedback. *Preprint*, arXiv:2403.18349.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2023. Alignment for honesty. *Preprint*, arXiv:2312.07000.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *Preprint*, arXiv:1809.09600.

Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. 2023. Cognitive mirage: A review of hallucinations in large language models. *Preprint*, arXiv:2309.06794.

Lei Yu, Meng Cao, Jackie Chi Kit Cheung, and Yue Dong. 2024. Mechanisms of non-factual hallucinations in language models. *Preprint*, arXiv:2403.18167.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi R. Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2023a. R-tuning: Teaching large language models to refuse unknown questions. *Preprint*, arXiv:2311.09677.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2023b. How language model hallucinations can snowball. *Preprint*, arXiv:2305.13534.

Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Lifeng Jin, Linfeng Song, Haitao Mi, and Helen Meng. 2024. Self-alignment for factuality: Mitigating hallucinations in llms via self-evaluation. *Preprint*, arXiv:2402.09267.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023c. Siren's song in the ai ocean: A survey on hallucination in large language models. *Preprint*, arXiv:2309.01219.

A Prompt for AI Feedback

Following Section 3.1, we design prompts for extraction, verification, and assessment tasks, as shown in Table 8, 9, and 10, respectively.

B Hyperparameters

We perform RLFH training with two different models: Qwen-2.5-7B-Instruct and Llama-3.1-8B-Instruct. We adopt different hyperparameter settings for the two models according to their own characteristics. For all experiments, we keep the same hyperparameter settings.

Table 4: Hyperparameter Settings for Different Models

Hyperparameter	**Qwen-2.5-7B**	**Llama-3.1-8B**
Actor Learning Rate	3e-7	5e-7
Critic Learning Rate	9e-6	9e-6
KL Coefficient	1e-2	5e-2
Train Batch Size	128	128
Rollout Batch Size	128	128
Episode	1	1
Advantage Estimator	GAE	GAE
GAE λ	0.95	0.95
Truthfulness Weight α	1	1
Informativeness Weight β	1.2	1.2
Informativeness ϵ	-0.9	-0.9
Informativeness μ	1.0	1.0
Verification Map Function f		
Correct	0.45	0.2
Hedged correct	0.35	0.1
Vague	-1.0	-1.8
Hedged wrong	-1.5	-2.0
Wrong	-1.7	-2.2
Informative Map Function G		
5	1.25	1.2
4	1.0	1.0
3	0.75	0.8
2	0.1	0.6
1	-0.2	-0.1

C Computation Resource

While our PPO-based implementation requires additional resources compared to FACT's DPO approach (Tian et al., 2023), including an additional value model and inference engines, the overall computational cost remains comparable for several reasons. First, FACT requires multiple response samples per prompt, whereas our method needs only one. Additionally, FACT relies on the FactScore pipeline (Min et al., 2023) typically run by a larger annotation model (Qwen2.5-72B in our experiments), while our method utilizes the model being trained (Qwen2.5-7B or Llama3.1-8B in our experiments) for self-assessment, significantly reducing annotation costs. In our experiments, each training run typically takes less than 1.5 hours using two 8-GPU nodes. Alternatively, training on a single 8-GPU node requires approximately 3 hours per run. Compared to traditional RLHF, our self-alignment approach eliminates the need for a separate reward

model, resulting in reduced computational requirements. Furthermore, while our implementation is based on PPO, more computationally efficient online reinforcement learning algorithms such as RLOO (Ahmadian et al., 2024), REINFORCE++ (Hu, 2025) or GRPO (Shao et al., 2024) could be adopted. These alternatives, which operate without a value model, would enable even more lightweight implementations of our approach.

D Factuality Metrics Validation

Following FactScore (Min et al., 2023), we validate the effectiveness of our hallucination evaluation. We conduct correlation analysis between Qwen2.5-72B-derived FactScore and human annotations across responses from three different models (InstructGPT, ChatGPT, and PerplexityAI). The validation results, as shown in Table 5, demonstrate strong alignment between our automated evaluation and human judgments. Specifically, the Pearson correlation coefficients (COEF) between Qwen2.5-72B and human annotations are notably high: 0.923 for InstGPT, 0.909 for ChatGPT, and 0.737 for PPLAI. Moreover, the error rates (ER) remain consistently low (0.041-0.098), which aligns with results from the original FactScore study. These strong correlations and low error rates across different model responses validate the reliability of our evaluation approach.

Table 5: Correlation analysis between human annotations and Qwen2.5-72B evaluations across different models.

Evaluator	InstGPT			ChatGPT			PPLAI		
	Score	COEF	ER	Score	COEF	ER	Score	COEF	ER
Human	0.428	–	–	0.583	–	–	0.804	–	–
Qwen2.5-72B	0.469	0.923	0.041	0.624	0.909	0.041	0.706	0.737	0.098

Statement Extraction Prompt

- Find every sentence containing object facts.
- Break sentences into atomic statements.
- Skip the sentences without statements.
- If there is no valid sentence, output "No statements".
- Do not output any explanation or other words.
- Strictly follow the output format shown in the example.

Here is an example:
Response
It is difficult to say which game has been released in more versions without more information, so I can only guess based on my training data.
Arthur's Magazine was likely started first. It was possibly founded in 1923 by Arthur K. Watson, a prominent publisher in the field of men's magazines.
First for Women, on the other hand, was not founded until 1989. It was created as a spin-off of Family Circle magazine, which was founded in 1957.

Statements
» Sentence 1: Arthur's Magazine was likely started first.
* Arthur's Magazine was likely started first.
» Sentence 2: It was possibly founded in 1923 by Arthur K. Watson, a prominent publisher in the field of men's magazines.
* Arthur's Magazine was possibly founded in 1923.
* Arthur's Magazine was founded by Arthur K. Watson.
* Arthur K. Watson is a prominent publisher in the field of men's magazines.
» Sentence 3: First for Women, on the other hand, was not founded until 1989.
* First for Women was not founded until 1989.
» Sentence 4: It was created as a spin-off of Family Circle magazine, which was founded in 1957.
* First for Women was created as a spin-off of Family Circle magazine.
* Family Circle magazine was founded in 1957.

And then comes your task:
Response
{response}

Statements

Figure 8: Template for extracting statements from the model responses.

Statement Verification Prompt

Choose from "Correct", "Vague" and "Wrong" for the verification of the statement.
- "Correct": The statement is supported by the materials.
- "Vague": Hard to determine the truthfulness of the statement based on the materials.
- "Wrong": The statement is negated by the materials.
Directly output the verification result without explanation.
Here is an example:

Materials
- First for Women is a women's magazine published by Bauer Media Group in the USA. The magazine was started in 1989. It is based in Englewood Cliffs, New Jersey. In 2011 the circulation of the magazine was 1,310,696 copies.
- Arthur's Magazine (1844–1846) was an American literary periodical published in Philadelphia in the 19th century. Edited by T.S. Arthur, it featured work by Edgar A. Poe, J.H. Ingraham, Sarah Josepha Hale, Thomas G. Spear, and others. In May 1846 it was merged into "Godey's Lady's Book".
- The correct answer for the question "Which magazine was started first Arthur's Magazine or First for Women" may be "Arthur's Magazine".
Statement
Arthur's Magazine was likely started first.
Verification
Correct

And then comes your task:
Materials
{materials}
Statement
{statement}
Verification

Figure 9: Template for verifying statement based on external material.

Statement Assessment Prompt

Evaluate the helpfulness of the statement:
- "5": The statement answer the question.
- "4": The statement provides crucial information.
- "3": The statement contains relevant facts.
- "2": The statement is about other supplementary facts.
- "1": The statement is useless or not relevant at all.
Directly output the evaluation result without explanation.

Here is an example:
Question
Which magazine was started first Arthur's Magazine founded by Arthur K. Watson or First for Women?
Response
It is difficult to say which game has been released in more versions without more information, so I can only guess based on my training data.
Arthur's Magazine was likely started first. It was possibly founded in 1923 by Arthur K. Watson, a prominent publisher in the field of men's magazines.
First for Women, on the other hand, was not founded until 1989. It was created as a spin-off of Family Circle magazine, which was founded in 1957.
Statement
Arthur's Magazine was possibly founded in 1923.
Evaluation
4

And then comes your task:
Question
{question}
Response
{response}
Statement
{statement}
Evaluation

Figure 10: Template for assessing statement importance based on original response.