MASK-DPO: GENERALIZABLE FINE-GRAINED FACTUALITY ALIGNMENT OF LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) exhibit hallucinations (*i.e.*, unfaithful or nonsensical information) when serving as AI assistants in various domains. Since hallucinations always come with truthful content in the LLM responses, previous factuality alignment methods that conduct response-level preference learning inevitably introduced noises during training. Therefore, this paper proposes a fine-grained factuality alignment method based on Direct Preference Optimization (DPO), called Mask-DPO. Incorporating sentence-level factuality as mask signals, Mask-DPO only learns from factually correct sentences in the preferred samples and prevents the penalty on factual contents in the not preferred samples, which resolves the ambiguity in the preference learning. Extensive experimental results demonstrate that Mask-DPO can significantly improve the factuality of LLMs responses to questions from both in-domain and out-of-domain datasets, although these questions and their corresponding topics are unseen during training. Only trained on the ANAH train set, the score of Llama3.1-8B-Instruct on the ANAH test set is improved from 49.19% to 77.53%, even surpassing the score of Llama3.1-70B-Instruct (53.44%), while its FactScore on the out-of-domain Biography dataset is also improved from 30.29% to 39.39%. We further study the generalization property of Mask-DPO using different training sample scaling strategies and find that scaling the number of topics in the dataset is more effective than the number of questions. We provide a hypothesis of what factual alignment is doing with LLMs, on the implication of this phenomenon, and conduct proof-of-concept experiments to verify it. We hope the method and the findings pave the way for future research on scaling factuality alignment.

1 INTRODUCTION

Large Language Models (LLMs) have demonstrated impressive performance across various tasks (Kamalloo et al., 2023; Sun et al., 2023; Chen et al., 2024). Even though, LLMs still face a concerning issue: *hallucination*, in which they generate plausible-sounding yet inaccurate or nonsensical information (Ji et al., 2022; Bang et al., 2023) in response to user queries, particularly those demanding extensive knowledge. The hallucination phenomenon has two main characteristics (Ji et al., 2024; Mishra et al., 2024) that significantly impede the real-world application of LLMs: 1) the unfaithful or nonsensical information always exists with truthful contents in the LLM responses, posing a challenge to accurately and effectively detect and mitigate them; 2) LLMs exhibit hallucinations across various domains and tasks, but it is difficult and impractical to thoroughly mitigate them by exhausting all real-world knowledge. The prevalence of these two issues underscores the need to develop effective and generalizable methods for fine-grained factuality alignment to reduce hallucinations in Large Language Models (LLMs).

Existing methods (Tian et al., 2023; Lin et al., 2024; Zhang et al., 2024b; Chen & Li, 2024) mainly apply preference learning-based method, especially Direct Preference Optimization (DPO) (Rafailov et al., 2024), to align the internal knowledge of LLMs to the facts. These approaches utilize the response-level factuality for constructing pairwise preference data for preference optimization, which aims to maximize the probability of preferred samples with higher factuality while minimizing the probability of non-preferred samples with lower factuality. However, since both factually correct and incorrect sentences are often mixed within a single response (Ji et al., 2024; Mishra et al., 2024), vanilla DPO inadvertently and inevitably encourages incorrect information in the preferred samples

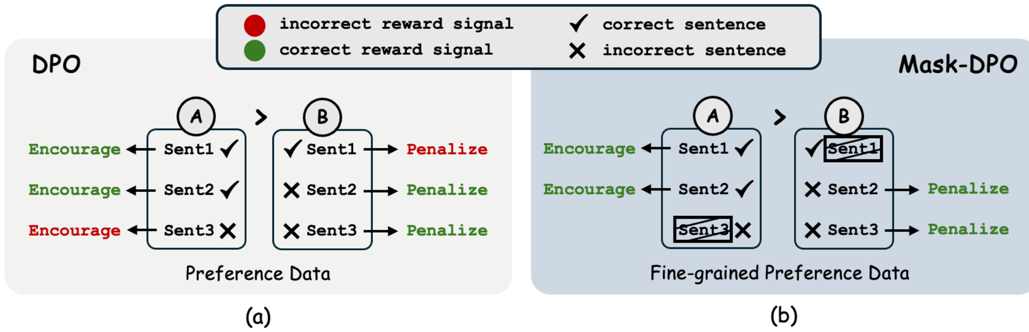

Figure 1: **Comparison between DPO and Mask-DPO.** Vanilla DPO (a) inadvertently encourages and penalizes all the content in the preferred and non-preferred samples, respectively, regardless of their correctness. Instead, Mask-DPO (b) incorporates sentence-level facticity into the mask signal, preventing incorrect reward signal, which resolves ambiguity in preference learning.

and penalizes truthful ones in the non-preferred samples (Figure 1(a)). Such ambiguity in the learning exists as long as the training samples are not completely wrong or correct, which is often the case, ultimately reducing the effectiveness of factuality alignment.

To address the issue, this paper proposes a fine-grained factuality alignment framework, named Mask-DPO. As shown in Figure 1, unlike vanilla DPO, Mask-DPO leverages sentence-level factuality information as a mask signal to resolve the ambiguity in preference learning. Specifically, in the preference data construction stage, Mask-DPO uses a fine-grained hallucination annotator (*e.g.*, ANAH-v2 (Gu et al., 2024)) to determine the factual correctness of each sentence, which will be used to guide the preference learning later. The responses that contain more and fewer factually correct sentences will be chosen as preferred and non-preferred samples, respectively, in a preference pair for learning. During the preference learning stage, the sentences that cause ambiguities in training, *i.e.*, incorrect sentences in the preferred samples and correct sentences in the not preferred samples, would be ignored, guided by the sentence-level mask annotated in the previous stage. Thus, Mask-DPO can avoid encouraging incorrect sentences when maximizing the probability of the preferred sample, and avoid penalizing correct sentences when minimizing the probability of the non-preferred sample.

Through extensive experiments, we demonstrate that Mask-DPO significantly enhances the factuality of LLMs and exhibits more effectiveness than DPO. Mask-DPO boosts the score of Llama3.1-8B-Instruct on the ANAH test set from 49.19% to 77.53%, surpassing the Llama3.1-70B-Instruct (53.44%) and DPO (68.44%), although the questions and topics in the test set are unseen during the alignment. Moreover, when testing the same model on the out-of-domain Biography dataset, Llama3.1-8B-Instruct aligned by Mask-DPO also exhibits excellent performance, whose FactScore is improved from 30.29% to 39.39%, reaching a level close to Llama3.1-70B-Instruct (40.47%).

We further study the scaling and generalization property of Mask-DPO, by scaling the training data through two dimensions: the number of different topics and the diversity of questions within the same range of topics. The experimental results show that scaling the number of topics is more effective for improving the factuality and generalization of models, in comparison to increasing the diversity of questions under the same topics, under the constraints of using the same amount of preference pairs. On the implication of this phenomenon, we hypothesize that each LLM learns a model-specific knowledge graph during training with the language modeling objective, where different topics and their corresponding information can be regarded as graph nodes, and the affinity between two nodes positively determines the probability of blurring their corresponding knowledge when the LLM responds to the questions related to either one of the topics.

Under this hypothesis, factual alignment on a topic essentially adjusts the affinity between the topic and its nearby nodes, which will also help to improve the factuality when the model answers questions of the nearby topics, even if they are not seen in the alignment. We conduct two proof-of-concept experiments to verify our hypothesis. The first one ablates different topic sampling strategies and shows that sampling the nearest topics to those in the test set based on the model-specific clustering can most effectively improve the factuality score on the test set. The comparison between the factuality of best-of-N responses of the model before and after factuality alignment further verifies

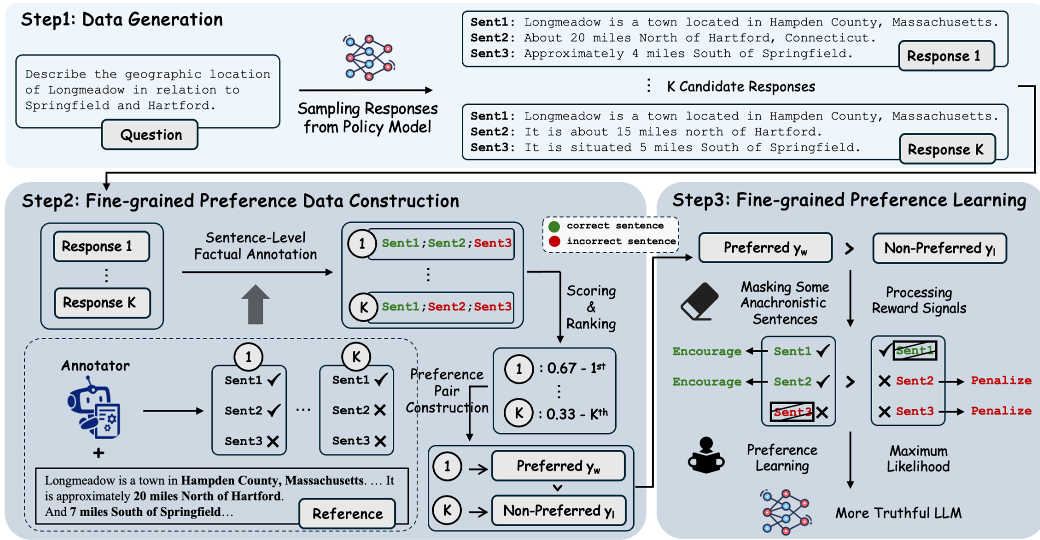

Figure 2: **The overview of Mask-DPO.** First, we sample K candidate responses for each question from the policy model. Then, we use a fine-grained hallucination annotator to perform a sentence-level factuality annotation on each response. We use the proportion of correct sentences out of the total number of sentences as the factuality score. We select the responses with the highest and lowest scores as preferred and non-preferred samples, respectively. Finally, we perform fine-grained factuality alignment on the policy model using such fine-grained preference data, where the reward signals to those anachronistic sentences, *i.e.*, incorrect sentences in the preferred samples and correct sentences in the non-preferred samples, would be ignored.

that the internal knowledge of LLMs is indeed more aligned to the facts after factuality alignment, which has not been observed in previous experiments of preference learning on other domains such as reasoning (Havrilla et al., 2024). We wish Mask-DPO and our findings could pave the way for future research on scaling factuality alignment of LLMs.

2 Mask-DPO

Mask-DPO is built on a preference-based reinforcement learning framework (§ 2.1). While previous methods optimize preferences at the response level which introduces noises, our approach employs dense supervision at the sentence level, and performs a fine-grained preference learning(§ 2.2), as shown in Fig. 2.

2.1 Preliminaries

Reinforcement Learning from Human Feedback (RLHF) is a powerful alignment method for fine-tuning language models to enhance their robustness, factuality, and safety (Ouyang et al., 2022). In practice, given a prompt x and a corresponding LLM response y, RLHF aims to maximize the following objective:

$$\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D}_p, y\sim\pi_\theta(\cdot|x)} \left[r(x,y) - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right], \tag{1}$$

where $\mathcal{D}_p$ represents the prompts dataset, π_θ is the policy model to be optimized, π_{ref} is the reference model used for regularizing π_θ with Kullback–Leibler divergence and β is a constant to control the degree of regularization. The reward model r reflects human preferences, which takes a prompt and the corresponding response as input and outputs a scalar value.

However, training a reward model and integrating it into the overall pipeline can be quite complex (Zheng et al., 2023). To avoid this, Rafailov et al. (2024) proposed Direct Preference Optimization (DPO), which directly optimizes the policy model using preference pairs. Given a prompt x, a preference pair that contains y_w and y_l sampled from the policy model π_θ, where y_w has higher quality than y_l. DPO aims to maximize the probability of the preferred sample y_w while minimizing the probability of the less desirable sample y_l. The optimization objective is formulated as:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right] \tag{2}$$

where $\mathcal{D}$ is the pair-wise preference dataset and σ is the sigmoid function.

Preferece learning-based alignment has been proven effective for factuality tuning in mitigating the hallucination of LLMs (Tian et al., 2023; Lin et al., 2024; Zhang et al., 2024b). However, the optimization objective of Eq. 2 may not be entirely suitable for knowledge-based tasks, where correct sentences and incorrect sentences are often coupled in a single response (Ji et al., 2024). Taking the preferred sample y_w as an example, we assume it has N sentences, *i.e.*$y_w = \{s_i\}_{i=1}^N$. The log of its generative probabilities in Eq. 2 can be expanded as:

$$\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} = \log\frac{\prod_{i=1}^N \pi_\theta(s_i|x, s_{1\sim i-1})}{\prod_{i=1}^N \pi_{\text{ref}}(s_i|x, s_{1\sim i-1})} = \sum_{i=1}^N \log\frac{\pi_\theta(s_i|x, s_{1\sim i-1})}{\pi_{\text{ref}}(s_i|x, s_{1\sim i-1})} \tag{3}$$

In such case, once there is a sentence s_i in y_w that contains incorrect facts, the generation probability of s_i would be maximized when maximizing the generation probability of y_w. Similarly, the generation probability of the correct sentence in y_l would be minimized when minimizing the y_l generation probability. This affects the effectiveness of factuality alignment.

2.2 Fine-grained Factuality Alignment

Fine-grained Preference Data Construction. We perform fine-grained factuality annotation using a sentence-level hallucination annotator (or reward model). Given the prompt x and its corresponding response y, which consist of N sentences $s = \{s_i\}_{i=1}^N$, the reward model $\mathcal{A}$ annotate each sentence and give the fine-grained feedback $a = \{a_i\}_{i=1}^N$, which can be formulated as:

$$a = \{a_i\}_{i=1}^N = \{\mathcal{A}(s_i, x)\}_{i=1}^N = \mathcal{A}(y, x) \tag{4}$$

Here, we let $a_i = 0$ represent that sentence s_i has no hallucination, and $a_i = 1$ represents that it has the hallucination.

Thanks to the fine-grained factuality annotation, we can now construct the preference dataset for alignment. First, for each prompt x, we perform top-k sampling to select K candidate responses $y = \{y^i\}_{i=0}^K$ from the policy model. For each candidate, we annotate it according to Eq. 4 and obtain $\{x, y^i, s^i, a^i\}_{i=0}^K$. Next, we compute the factuality score for each candidate, which we define as the fraction of correct facts: $\frac{\sum_{i=0}^N (1-a_i)}{N}$, where N is the number of sentences in candidate y^i. Then, we construct a preference dataset based on factuality scores, where we assign y^i with higher scores to y_w and the lower ones to y_l. Finally, we filter the data to filter out preference pairs where y_w and y_l are the same, y_w does not contain correct facts, and y_l does not contain incorrect facts.

Fine-grained Preference Learning. Using the constructed fine-grained preference dataset $\mathcal{D} = \{x, y_w, s_w, a_w, y_l, s_l, a_l\}$, Mask-DPO forces the model to only learn from the correct facts in the preferred examples y_w and the incorrect contents in the un-preferred examples y_l. Based on Eq. 3, we designed a masking scheme applied to the DPO algorithm as:

$$\begin{aligned}\mathcal{M}_w(x, s_w, a_w) &= \sum_i \mathbb{I}(a_i = 0)\log\frac{\pi_\theta(s_i|x, s_{1\sim i-1})}{\pi_{\text{ref}}(s_i|x, s_{1\sim i-1})}\\ \mathcal{M}_l(x, s_w, a_w) &= \sum_j \mathbb{I}(a_j = 1)\log\frac{\pi_\theta(s_j|x, s_{j\sim j-1})}{\pi_{\text{ref}}(s_j|x, s_{1\sim j-1})}\end{aligned} \tag{5}$$

where $\mathcal{M}_w$ is the masked KL divergence for y_w while $\mathcal{M}_l$ for y_l. Combining Eq. 2 and Eq. 5, we formulate the optimization objective of fine-grained factuality tuning:

$$\mathcal{L}_{\text{Mask-DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,s_w,a_w,s_l,a_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\mathcal{M}_w(x, s_w, a_w) - \beta\mathcal{M}_l(x, s_l, a_l)\right)\right] \tag{6}$$

Compared with the original DPO optimization objective, our Mask-DPO does not maximize the probability of incorrect facts in y_w, nor minimize the probability of correct facts in y_l.

Table 1: **Evaluation results for the open-source models, FactTune, and our Mask-DPO.** Here, ANAH and Biography represent the in-domain and out-of-domain test sets, respectively. ANAH-v2 and FactScore represent the corresponding evaluation strategies. For each evaluation strategy, we report the number of correct facts (# Correct.), the number of inaccuracies facts (# Inc.), and the factuality score (% Score), *i.e.* the proportion of correct facts out of the total number of facts. **Bold** and underlined represent the best and second best performance, respectively.

Model	ANAH (in-domain)						Biography (out-of-domain)		
	ANAH-v2 Evaluator			*Factscore Evaluator*			*Factscore Evaluator*		
	# Cor.	# Inc.	% Score ↑	# Cor.	# Inc.	% Score ↑	# Cor.	# Inc.	% Score ↑
Qwen2-7B	450	872	34.03	5.19	30.60	15.57	14.46	39.27	27.28
Qwen2-72B	538	690	43.81	6.45	27.97	20.76	20.13	40.09	34.06
Gemma2-9B	635	837	43.13	4.51	25.92	18.29	15.53	29	29.52
Gemma2-27B	889	1110	44.47	7.81	35.65	20.14	19.27	36.59	29.99
Yi1.5-6B	440	1143	27.79	6.01	44.48	11.92	10.84	59.65	15.81
Yi1.5-9B	483	1136	29.83	5.06	39.79	10.42	10.63	62.67	15.33
Yi1.5-34B	535	911	36.99	5.25	36.96	13.27	17.01	52.76	25.49
Llama3.1-8B	461	476	49.19	5.95	21.29	19.43	16.83	31.57	30.29
Llama3.1-70B	520	453	53.44	6.81	22.17	21.92	23.58	31.59	**40.47**
FactTune	657	499	56.83	7.45	23.08	22.67	15.93	23.97	37.97
Ours	547.6	161.4	**77.53**	5.9	18.59	**25.56**	12.16	15.24	39.39

3 EXPERIMENT

3.1 EXPERIMENT SETUP

Implementation. In our experimental framework, we adopt the Llama3.1-8B-Instruct (Dubey et al., 2024) as the base model and the ANAH-v2 (Gu et al., 2024) as the fine-grained reward model to annotate factuality. All the experiments are conducted using the same setting for a fair comparison. Further implementation details can be found in Appendix A.

Dataset. For the in-domain data, we use a subset of the ANAH-v2 (Gu et al., 2024) data as the test set, containing 177 questions, and another 8046 questions as the training set. ANAH is organized along two dimensions: topics and questions, and the training set has no overlap with the test set in both dimensions. For the out-of-domain data, we use a subset of Biography (Min et al., 2023) as the test set, which has 183 questions about biography generation and has no overlap with our training set.

Evaluation. To evaluate each generated response, we use counts of correct and incorrect facts computed by ANAH-v2 (Gu et al., 2024) and FactScore (Min et al., 2023) as the evaluation metrics. ANAH-v2, which was trained specifically on our in-domain data for the hallucination annotation task, would annotate each sentence in the response with hallucinations, and we used the ratio of non-hallucinated sentences to the total number of sentences as the final score. Since ANAH-v2 has been used in preference data construction, to avoid reward hacking, we also use FactScore for evaluation, which calculates the number of correct facts and reports its proportion. In addition, considering that ANAH-v2 has not yet been trained in the task of generating biographies, and to ensure the reliability of the results, we only use FactScore as an evaluation strategy when evaluating out-of-domain data.

Baseline. We construct the baseline in two ways. First, we compare Mask-DPO with many open-source models, including Qwen2 (7B & 72B) (Yang et al., 2024), Gemma2 (9B & 27B) (Team et al., 2024), Yi1.5 (6B & 9B & 34B) (Young et al., 2024), and Llama3.1 (8B & 70B) (Dubey et al., 2024). Note that all open-source models used here are post-trained versions. In addition, we also compare FactTune (Tian et al., 2023), a hallucination mitigation method that is also based on DPO. We apply it to the same base model.

3.2 OVERALL RESULTS

Evaluation on In-Domain Data. The first six columns of Tabel 1 illustrate the performance of the open source models, FactTune, and our Mask-DPO on the in-domain data. Under the ANAH-v2 evaluation strategy, Mask-DPO achieves the highest factuality score (77.53%), surpassing the best

Table 2: **Ablation study for the masking scheme applied to the DPO.** Here, "w/ mask" means that the factual alignment algorithm is DPO with mask, which is the default setting of Mask-DPO. "w/o mask" means that the algorithm is the vanilla DPO.

Training Setting	ANAH-v2			FactScore		
	# Correct	# Incorrect	% Score ↑	# Correct	# Incorrect	% Score ↑
w/o mask	446.5	206	68.44	5.71	18.98	23.43
w/ mask	547.6	161.4	**77.53**	5.9	18.59	**25.56**

Table 3: **Ablation study for the sampling strategy in preference dataset construction.** Here, "Llama" means that the data used to construct preference pairs is sampled from the policy model (Llama3.1-8B-Instruct), which is the default setting of Mask-DPO. "InternLM" means that the data is sampled from InternLM2-7B-Chat-SFT. And "Llama+Doc" means that the data is sampled from the policy model, but the sampling strategies for preferred and un-preferred samples are different. When generating preferred samples, the reference document corresponding to the question is added to the prompt, but not to the un-preferred ones.

Sampling Setting	ANAH-v2			FactScore		
	# Correct	# Incorrect	% Score ↑	# Correct	# Incorrect	% Score ↑
InternLM	419.25	296.75	59.43	5.19	16.59	21.33
Llama+Doc	877	1347	39.43	7.98	33.07	16.16
Llama	547.6	161.4	**77.53**	5.0	18.59	**25.56**

open-source model Llama3.1-70B-Instruct (53.44%) and factuality alignment method FactTune (56.83%). In addition, under the FactScore evaluation strategy, the results also show the same trend, where Mask-DPO reaches the highest factuality score (25.56%), surpassing the best open-source model Llama3.1-70B-Instruct (21.92%) and factuality alignment method FactTune (22.67%). These results also show that our reward model has not been overly and seriously hacked.

Tabel 1 also provides the factuality levels of existing open-source models, among which the Llama3.1 series achieves the highest factuality level, while the Yi1.5 series has the lowest level. Moreover, as the scale of model parameters increases, the factuality level of the model also increases, which is consistent with the observations of Gu et al. (2024).

Evaluation on Out-of-Domain Data. The last three columns of Tabel 1 illustrate the performance of the open source model, FactTune, and our Mask-DPO on the out-of-domain data. Without using the corresponding training set, Mask-DPO still exhibits excellent results, improving the factuality score of Llama3.1-8B-Instruct on the Biography dataset from 30.29% to 39.39%, surpassing the factuality alignment method FactTune (37.97%) and reaching a level close to the best open-source model Llama3.1-70B-Instruct (40.47%).

Moreover, the trends of the factuality levels between different series of open-source models and between different parameter amounts in the same series are consistent with those in Table 1, further illustrating the reliability of our results.

3.3 Ablation Study

Impact of Mask Scheme. To verify the effectiveness of the masking scheme (introduced in § 2.2), we compare the performance of the models trained with and without the masking scheme. As shown in Tabel 2, the scheme with the mask is significantly better than the scheme without the mask in both evaluation strategies, which demonstrates the effectiveness of Mask-DPO.

It is worth noting that FactTune (the second to last row in Table 1) also uses the vanilla DPO for factuality alignment, but its final performance is not as good as the vanilla DPO in this experiment (the first row in Table 2). The main difference between them is that FactTune uses FactScore to construct preference data, while the vanilla DPO in this experiment uses ANAH-v2. Because FactScore and ANAH-v2 are also our two evaluation models, this result is shown in both evaluation strategies, which rules out the possibility of rewarding hackers. Therefore, we believe that the reward model used to construct preference data will have an obvious impact on the effectiveness of factuality alignment.

Table 4: **Evaluation for the efficiency and the generalization effects of different training sample scaling strategies.** Here, "Topic" means scaling from the number of different topics, and "Question" means scaling from the number of different questions under the same topic. " Topic" denotes the number of topics used for training, and " Question" denotes the number of questions under the same topic used for training. We also report changes in scores (in parentheses) when scaling the data.

Dimension	# Topic	# Question	Scale	# Correct	# Incorrect	% Score ↑
Topic	894	3	1/3	501.75	270.25	65.04
	1788	3	2/3	518.00	206.25	71.56 (+6.53)
	2682	3	1	547.60	161.40	77.53 (+5.97)
Question	2682	1	1/3	486.00	218.00	69.03
	2682	2	2/3	444.00	164.00	73.03 (+4.00)
	2682	3	1	547.60	161.40	77.53 (+4.50)

Impact of Sampling Strategy. To assess the impact of the data sampling strategy (introduced in § 2.2), we compare the performance of the models trained with different data sampling strategies in Tabel 3. In our default setting, we sample the data from the policy model (Llama) to construct the factual preference dataset. We also test two other different sampling strategies, sampling from the non-policy model (InternLM) and sampling from the policy model but using different contexts for preferred and non-preferred samples (Llama + Doc).

As shown in Tabel 3, the performance of sampling from the non-policy model is far less effective than sampling from the policy model, even though it can improve on the base model. Notably, for settings that sample from the policy model but use different contexts for preferred and non-preferred samples, the scores after factuality alignment even drop from the base model. Intuitively, the quality of preference pairs constructed in this setting should be higher due to the presence of corresponding reference documents in the prompt while generating the preferred samples. This may be the case because the prompts used to generate the preferred and non-preferred samples are not the same, resulting in a negative impact on the factuality alignment although the quality of the preference pairs is higher. This phenomenon is also observed by Lin et al. (2024).

4 Analysis of Data Scaling

We further analyze the efficiency and the generalization effects of different training sample scaling strategies of Mask-DPO (§ 4.1). To explain the observed phenomenon of generalization, we provide a hypothesis that discusses how factuality alignment implicitly impacts the internal knowledge structure in LLMs (§ 4.2), and then briefly design a series of proof-of-concept experiments to verify it. (§ 4.3).

4.1 Analysis

In our training set, data are organized along two dimensions: topics and questions, where each question can be categorized into a topic, which is usually the entities the question asks about. Both topics and questions in the training set have no overlap with that of the test set. We analyze the impact of scaling the training data through these two dimensions, including the number of different topics and the diversity of questions under the same topic, respectively.

In the default setting of Mask-DPO, we use 2682 topics for factuality alignment, where each topic contains 3 questions. In the scaling of the number of topics, we use one-third, two-thirds, and the full number of topics for alignment. In the question dimension, we use the same scaling strategy, *i.e.*, only use one, two, or three questions per topic for alignment. Since we have demonstrated the consistency of ANAH-v2 and FactScore in terms of results in the previous section (§ 3), for resource considerations, we only use ANAH-v2 for the evaluation afterward.

As shown in Table 4, we observe that scaling the number of topics is more effective than increasing the diversity of questions. When scaling the same amount of data, scaling the topic leads to larger increases (6.53 and 5.97) compared to scaling the question (4.00 and 4.50). Moreover, the scores for the "scaling number of questions" setting are consistently higher than those for the "scaling number of topics" with the same amount of training data. Since the former consistently covers the full amount of topics, it suggests that the diversity of topics is more important than the diversity of questions.

4.2 Hypothesis

The experimental results in previous sections indicate that after factuality alignment with questions from a group of topics, the LLM will become more factually correct when answering questions from another group of topics. On the implication of this phenomenon, we hypothesize that the factuality alignment essentially adjusts a model-specific graph consisting of topics as nodes represented by the LLM, where preference learning on some of the topics (nodes in the graph) will be propagated to other nodes through the graph.

Specifically, consider a graph $G_\theta = (V, E)$, where V is the set of nodes, each representing a distinct topic and their corresponding knowledge, and $E \subseteq V \times V$ is the set of edges, representing relationships between topics. The edges are defined by the affinity between these topics, which are learned specifically for each model through the language modeling objective in training (especially pre-training). Thus, the hallucination can be interpreted as the blurring of knowledge between a node and other nodes. For any two topics $u, v \in V$, given the user query x, the hallucination phenomenon can be formulated as:

$$h(x, v \to u) \implies s_i = \pi_\theta\left(\cdot \mid x, s_{1\sim i-1}\right), \quad x, s_{1\sim i-1} \in v,\, s_i \in u, \tag{7}$$

where $s_{1\sim i-1}$ represents the first $i-1$ sentences in the reply of LLM, related to topic v, and s_i represents the i^{th} sentence, which erroneously shifts to information of topic u due to the boundary blurring between topics u and v.

Let an embedding function $\phi : V \to \mathbb{R}^d$ maps each topic $v \in V$ to a high-dimensional vector $\phi(v)$ in the embedding space $\mathbb{R}^d$. Due to the duality of the graph structure, the relationship between v and u can be quantified by a distance metric D, which we hypothesize can be approximated by the probability of its hallucination phenomenon:

$$D(\phi(u), \phi(v)) \propto \mathcal{P}(h(x, u \to v)) \approx \mathcal{P}(h(x, v \to u)) = \pi_\theta\left(s_i \mid x, s_{1\sim i-1}\right), x, s_{1\sim i-1} \in v,\, s_i \in u, \tag{8}$$

We hypothesize that more related topics will have smaller distances in the embedding space represented by LLMs, and thus are more likely to exhibit hallucination. This also explains that s_i often behaves plausibly and is hard to detect (Ji et al., 2022) due to the proximity of u and v.

With this formulation, factuality alignment using Eq. 2 and Eq. 3 on topic v can be considered as decreasing the probability of $h(x, v \to u)$. Due to the duality of the affinity between nodes (Eq. 8), nearby topics like u are implicitly adjusted together, which explains the generalization phenomenon. Moreover, Mask-DPO introduces a more accurate adjustment on each topic using Eq. 5 and Eq. 6, which makes the factuality alignment more effective.

4.3 Proof of concept

If the model-specific graph does exist, and our interpretation of factuality alignment on the graph is correct, then there will be two corollaries: 1) when factuality alignment is performed on a topic, topics closer to it will be more affected, as evidenced by the factuality score of that topic. 2) factuality alignment does not simply change the generation probability distribution or the style of certain content but has modified the internal knowledge structure of LLMs, which can be observed through best-of-N sampling.

Proof of Corollaries (1). To assess the impact between topics, we design an experiment based on topic clustering. We use different embedding strategies to get the high-dimensional vector of each topic, including embeddings using the policy model itself, embeddings from other models, and embeddings from specialized word embedding models. Then, we cluster the topics in the training set using the topic vectors in the test set as cluster centers, and split the training set into two groups, namely the far group and the near group, according to the distance from the corresponding topic vector to the cluster center. We also use a random strategy to select two groups as baselines. Finally, we use the far group and the near group for factual alignment under the same settings.

We conduct experiments using Llama3.1-8B-Instruct (Dubey et al., 2024) and InternLM2-7B-SFT-Chat (Cai et al., 2024) as the base (policy) models and use ANAH-v2 for evaluation.

Table 5: **Evaluation for the effects of different clustering strategies for preference data construction.** Here, "InternLM2-7B", "Llama3.1-8B" and "OpenAI" denotes InternLM2-7B-Chat-SFT, Llama3.1-8B-Instruct and Text-Embedding-3-Large, respectively. And "Random" means randomly selecting samples. "Distance" represents the distance between the training set and the cluster center. "Diff" represents the difference between the score at near distance settings and the one at far.

Base Model	Embedding Model	Distance	# Correct	# Incorrect	% Score ↑	Diff ↑
InternLM2-7B	InternLM2-7B	far	493.10	267.60	64.99	**2.43**
		near	456.40	222.30	**67.42**	
	Llama3.1-8B	far	454.80	252.20	64.32	0.65
		near	487.00	263.00	64.97	
	OpenAI	far	487.64	256.09	65.59	1.25
		near	489.50	242.70	66.84	
	Random	-	453.60	247.40	64.78	0.77
		-	521.70	275.90	65.55	
Llama3.1-8B	InternLM2-7B	far	455.50	206.50	68.97	0.22
		near	551.75	280.75	69.19	
	Llama3.1-8B	far	568.00	265.75	68.94	**4.49**
		near	500.00	182.33	**73.43**	
	OpenAI	far	761.25	317.00	71.32	0.85
		near	616.60	253.60	72.17	
	Random	-	566.00	219.00	70.55	0.56
		-	454.00	182.33	71.11	

Table 6: **Evaluation for the performance under the setting of best-of-N.** Here, "Baseline" represents the base model Llama3.1-8N-Instruct and "Mask-DPO" represents the aligned model. [1]

Model	Best of 1	Best of 16	Best of 32	Best of 64	Best of 128	Best of 256
Baseline	47.79	73.38	76.30	76.93	79.42	80.98
Mask-DPO	70.67	89.8	90.53	92.81	93.47	94.09

As shown in Table 5, regardless of the embedding strategy used, the factual score of alignment using the near group is higher than that using the far group. This indicates when factuality alignment is performed on a node, it affects other nodes, and the effect becomes smaller as the distance between nodes increases. Furthermore, when the base model and the embedding model are the same, the score difference between the far and near groups is the largest, *i.e.*, 2.43 for InternLM2-7B and 4.49 for Llama3.1-8B, and the near group obtains the best results (67.2% for InternLM2-7B and 73.43% for Llama3.1-8B), while the score differences for the other strategies are close to those for the random strategy and the near group chose by other strategies cannot obtain the highest results. This suggests that the graph-like knowledge structure is model-specific.

Proof of Corollaries (2). We further test the best-of-N performance of the Llama3.1-8B-Instruct (Dubey et al., 2024) before and after training, using ANAH-v2 for evaluation. Specifically, we perform top-k (k = 40) sampling in inference time, and chose the most factual response from N candidates as the final response for each question.

As shown in Table 6, the best-of-N scores for both settings rise in parallel as the number of samples increases. However, the difference in scores before and after factuality alignment is always present and does not gradually converge as the number of samples increases. This is inconsistent with what Havrilla et al. (2024) observed in the math reasoning task, where the scores before and after training would converge as the number of samples increased, which means that the training only changes the probability distribution that generates the correct answer. Such a phenomenon proves our corollaries that factuality alignment has an essential effect on the knowledge structure within LLMs and does not simply change the generation probability distribution of certain content.

[1] The results for best-of-1 do not match those in Table 1 because the inference setting here is top-k=40, whereas in Table 1 it is top-k=1.

5 Related Work

Reinforcement Learning from Human Feedback. In order to unlock the capabilities of LLMs and align them with human preferences, a bunch of work about alignment has been proposed. Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022; Bai et al., 2022; Rafailov et al., 2024; Yuan et al., 2024) is the typical alignment method, a representative work is Proximal Policy Optimization (PPO) (Schulman et al., 2017). PPO uses paired data based on human preferences to train a reward model and then uses the signals provided by that reward model to guide the optimization of the policy model, which is complex. For simplicity, Direct Preference Optimization (DPO) (Rafailov et al., 2024) utilizes pairwise preference data directly for model optimization. Since using humans for preference data construction is costly (Min et al., 2023), Reinforcement Learning from AI Feedback (RLAIF) (Lee et al., 2023) is proposed, which constructs preference data using AI (*e.g.* advanced LLMs, specialized evaluator models, and agents system) feedback. Our approach can be regarded as a fine-grained RLAIF method for hallucination mitigation, which uses feedback from the hallucination annotator as the reward signal and we improve DPO for more effective factuality alignment.

Hallucination Mitigation. Considering the harm of hallucinations, researchers have explored various techniques for enhancing the factuality of LLMs. In the inference stage, various techniques such as trustworthy decoding strategies (Rebuffel et al., 2022; Chuang et al., 2023; Shi et al., 2023; Li et al., 2023), multi-agent methods (Du et al., 2023), and variants of the Chain-of-Thought approach involving verification or reflection (Dhuliawala et al., 2023; Lei et al., 2023; Ji et al., 2023b; Wang et al., 2023; Gou et al., 2023) are proposed to improve the reliability of the output from LLMs. For the training stage, multi-task learning (Garg et al., 2019; Weng et al., 2020), model editing (Daheim et al., 2023; Ji et al., 2023a), and factuality alignment (Wu et al., 2023; Tian et al., 2023; Lin et al., 2024; Zhang et al., 2024b; Wen et al., 2024; Xu et al., 2024; Chen & Li, 2024) are attempted. Some of them (Wu et al., 2023; Wen et al., 2024) perform fine-grained RLHF using PPO, while our approach is based on the more efficient and cost-friendly DPO. Most similar to ours, some approaches attempt to mitigate hallucination using DPO, and their main difference lies in the different ways of constructing preference data. For example, Lin et al. (2024) uses an external estimator or annotator to construct factual preference data, while Chen & Li (2024), Zhang et al. (2024b) rely on the model itself, and Tian et al. (2023) has both solutions. However, all of them only use the factuality of the entire response as the sparse reward signal, making factuality alignment less effective. In contrast, our approach implements fine-grained factuality alignment using sentence-level reward signals.

Knowledge Editing. The primary cause of the hallucination is the fragility of parametric knowledge in LLMs (Wang et al., 2024a). To keep knowledge truthfulness and reliability, many works attempt to edit it, including knowledge and representation editing. Knowledge editing (Zhang et al., 2024a; Wang et al., 2024b; McAleese et al., 2024) focuses on selectively adjusting model parameters that store specific pieces of knowledge, while representation editing (Zou et al., 2023; Wu et al., 2024) modifies how the model conceptualizes knowledge to update the information retained within LLMs. Unlike them, our approach does not explicitly edit the knowledge-related weights or representations but we hypothesize factuality alignment implicitly modifies the knowledge structure inside the model.

6 Conclusion

In this paper, we introduce a fine-grained factuality alignment method, Mask-DPO. Incorporating sentence-level factuality as reward signals, Mask-DPO accurately masks the encouraging signal of incorrect sentences in the preferred samples and the punishment signal from factual contents in the non-preferred samples, thus improving the effectiveness of learning. Extensive experimental results demonstrate that Mask-DPO significantly increases the factuality of LLMs in both in-domain and out-of-domain evaluation. The study on the generalization property of Mask-DPO using different training sample scaling strategies reveals that the diversity of topics is more important to the generalization than the number of questions. On the implication of this observation, we hypothesize that LLM learns a model-specific graph-like knowledge structure in training, which propagates the impact of factuality alignment based on the distance between topics, where nearby topics will also be implicitly aligned, even if they are not seen during alignment. We conduct proof-of-concept experiments and show that factuality alignment indeed essentially aligns the internal knowledge of LLMs. We hope our study could pave the way for future research on scaling the factuality alignment of LLMs.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024.

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

Zehui Chen, Kuikun Liu, Qiuchen Wang, Jiangning Liu, Wenwei Zhang, Kai Chen, and Feng Zhao. Mindsearch: Mimicking human minds elicits deep ai searcher. *arXiv preprint arXiv:2407.20183*, 2024.

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

LMDeploy Contributors. Lmdeploy: A toolkit for compressing, deploying, and serving llm. `https://github.com/InternLM/lmdeploy`, 2023a.

XTuner Contributors. Xtuner: A toolkit for efficiently fine-tuning llm. `https://github.com/InternLM/xtuner`, 2023b.

Nico Daheim, Nouha Dziri, Mrinmaya Sachan, Iryna Gurevych, and Edoardo M Ponti. Elastic weight removal for faithful and abstractive dialogue generation. *arXiv preprint arXiv:2303.17574*, 2023.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*, 2023.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. 5 2023. URL `http://arxiv.org/abs/2305.14325`.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Sarthak Garg, Stephan Peitz, Udhyakumar Nallasamy, and Matthias Paulik. Jointly learning to align and translate with transformer models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4453–4462, 2019.

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

Yuzhe Gu, Ziwei Ji, Wenwei Zhang, Chengqi Lyu, Dahua Lin, and Kai Chen. Anah-v2: Scaling analytical hallucination annotation of large language models. *arXiv preprint arXiv:2407.04693*, 2024.

Alex Havrilla, Yuqing Du, Sharath Chandra Raparthy, Christoforos Nalmpantis, Jane Dwivedi-Yu, Maksym Zhuravinskyi, Eric Hambro, Sainbayar Sukhbaatar, and Roberta Raileanu. Teaching large language models to reason with reinforcement learning. *arXiv preprint arXiv:2403.04642*, 2024.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 2022.

Ziwei Ji, Zihan Liu, Nayeon Lee, Tiezheng Yu, Bryan Wilie, Min Zeng, and Pascale Fung. RHO: Reducing hallucination in open-domain dialogues with knowledge grounding. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 4504–4522, Toronto, Canada, July 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.275. URL https://aclanthology.org/2023.findings-acl.275.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. Towards mitigating hallucination in large language models via self-reflection. *EMNLP Findings*, 2023b.

Ziwei Ji, Yuzhe Gu, Wenwei Zhang, Chengqi Lyu, Dahua Lin, and Kai Chen. Anah: Analytical annotation of hallucinations in large language models. *arXiv preprint arXiv:2405.20315*, 2024.

Ehsan Kamalloo, Nouha Dziri, Charles L. A. Clarke, and Davood Rafiei. Evaluating open-domain question answering in the era of large language models. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 5591–5606. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.307. URL https://doi.org/10.18653/v1/2023.acl-long.307.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, et al. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023.

Deren Lei, Yaxi Li, Mingyu Wang, Vincent Yun, Emily Ching, Eslam Kamal, et al. Chain of natural language inference for reducing large language model ungrounded hallucinations. *arXiv preprint arXiv:2310.03951*, 2023.

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

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

Nat McAleese, Rai Michael Pokorny, Juan Felipe Ceron Uribe, Evgenia Nitishinskaya, Maja Trebacz, and Jan Leike. Llm critics help catch llm bugs. *arXiv preprint arXiv:2407.00215*, 2024.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *EMNLP*, 2023. URL https://arxiv.org/abs/2305.14251.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. Fine-grained hallucination detection and editing for language models. *arXiv preprint arXiv:2401.06855*, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

Clément Rebuffel, Marco Roberti, Laure Soulier, Geoffrey Scoutheeten, Rossella Cancelliere, and Patrick Gallinari. Controlling hallucinations at word level in data-to-text generation. *Data Mining and Knowledge Discovery*, 36(1):318–354, 2022.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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

Kai Sun, Yifan Ethan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. Head-to-tail: How knowledgeable are large language models (llm)? aka will llms replace knowledge graphs? *arXiv preprint arXiv:2308.10168*, 2023.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

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

Mengru Wang, Yunzhi Yao, Ziwen Xu, Shuofei Qiao, Shumin Deng, Peng Wang, Xiang Chen, Jia-Chen Gu, Yong Jiang, Pengjun Xie, et al. Knowledge mechanisms in large language models: A survey and perspective. *arXiv preprint arXiv:2407.15017*, 2024a.

Xiaohan Wang, Shengyu Mao, Ningyu Zhang, Shumin Deng, Yunzhi Yao, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. Editing conceptual knowledge for large language models. *arXiv preprint arXiv:2403.06259*, 2024b.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. Unleashing cognitive synergy in large language models: A task-solving agent through multi-persona selfcollaboration. *arXiv preprint arXiv:2307.05300*, 1(2):3, 2023.

Xueru Wen, Xinyu Lu, Xinyan Guan, Yaojie Lu, Hongyu Lin, Ben He, Xianpei Han, and Le Sun. On-policy fine-grained knowledge feedback for hallucination mitigation. *arXiv preprint arXiv:2406.12221*, 2024.

Rongxiang Weng, Heng Yu, Xiangpeng Wei, and Weihua Luo. Towards enhancing faithfulness for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2675–2684, 2020.

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

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D Manning, and Christopher Potts. Reft: Representation finetuning for language models. *arXiv preprint arXiv:2404.03592*, 2024.

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

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*, 2024.

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, et al. A comprehensive study of knowledge editing for large language models. *arXiv preprint arXiv:2401.01286*, 2024a.

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

Rui Zheng, Shihan Dou, Songyang Gao, Yuan Hua, Wei Shen, Binghai Wang, Yan Liu, Senjie Jin, Qin Liu, Yuhao Zhou, et al. Secrets of rlhf in large language models part i: Ppo. *arXiv preprint arXiv:2307.04964*, 2023.

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

A Implementation Details

In our experimental framework, we adopt the Llama3.1-8B-Instruct (Dubey et al., 2024) as the base model and the ANAH-v2 (Gu et al., 2024) as the fine-grained reward model to annotate factuality.

For preference data construction, we generate 32 candidate responses per question and selected the ones with the highest and lowest scores to form preference pairs after fine-grained factuality estimation. The decoding strategy involves the top-k (k = 40) sampling with a temperature of 0.8.

For fine-grained factuality tuning, we train the base model with the following settings and hyper-parameters: the epoch is 3, the learning rate is 5e-6, the batch size is 64, and the AdamW optimizer is the cosine annealing learning rate scheduler.

Our model is trained on 8 NVIDIA A100 GPUs. The training and inference frameworks are Xtuner (Contributors, 2023b) and LMDeploy (Contributors, 2023a) respectively.

B Prompt

For hallucination annotation in fine-grained preference data construction, we use the prompt from Gu et al. (2024), including factual existence judgment, reference information extraction, and hallucination-type judgment.

C Limitation

Although this study presents a novel and effective fine-grained factuality alignment framework, there are some limitations. Despite Mask-DPO's performance in fine hallucination relief is remarkable, it remains somewhat exaggerated. Because the model we used to construct the preference data is the same as the one used for the review, this could pose a potential risk of reward hacking. Although we introduced additional evaluation models to show that this hacking phenomenon is not significant, it still cannot be completely avoided. Furthermore, we only perform evaluation on ANAH and Biography. However, these datasets might not encompass the full spectrum of real-world scenarios where hallucinations pose a problem. Lastly, this work primarily uses Llama3.1-8B-Instruct and InternLM2-7B-Chat-SFT as the backbone of the base model. Other different underlying models and different numbers of parameters are not explored.