GRATH: Gradual Self-Truthifying for Large Language Models

Weixin Chen [1] **Dawn Song** [2] **Bo Li** [1 3]

Abstract

Truthfulness is paramount for large language models (LLMs) as they are increasingly deployed in real-world applications. However, existing LLMs still struggle with generating truthful content, as evidenced by their modest performance on benchmarks like TruthfulQA. To address this issue, we propose GRAdual self-truTHifying (GRATH), a novel post-processing method to enhance truthfulness of LLMs. GRATH utilizes out-of-domain question prompts to generate pairwise truthfulness training data with each pair containing a question and its correct and incorrect answers, and then optimizes the model via direct preference optimization (DPO) to learn from the truthfulness difference between answer pairs. GRATH iteratively refines truthfulness data and updates the model, leading to a gradual improvement in model truthfulness in a self-supervised manner. Empirically, we evaluate GRATH using different 7B-LLMs and compare with LLMs with similar or even larger sizes on benchmark datasets. Our results show that GRATH effectively improves LLMs' truthfulness without compromising other core capabilities. Notably, GRATH achieves state-of-the-art performance on TruthfulQA, with MC1 accuracy of 54.71% and MC2 accuracy of 69.10%, which even surpass those on 70B-LLMs. The code is available at https://github.com/chenweixin107/GRATH.

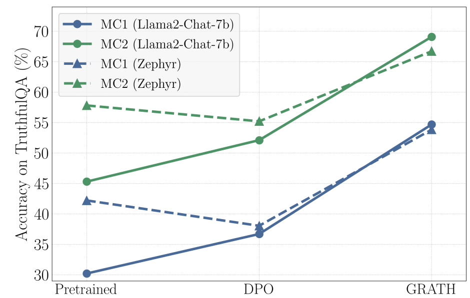

Figure 1. Accuracy of pretrained models, DPO, and GRATH on TruthfulQA's MC1 and MC2 tasks. We evaluate DPO and GRATH on two pretrained models, Llama2-Chat-7B and Zephyr, based on an OOD training dataset—ARC-Challenge. GRATH effectively improves MC1 and MC2 accuracy of Llama2-Chat-7B(Zephyr) by 24.5%(11.6%) and 23.8%(8.9%). DPO enhances MC1 and MC2 accuracy of Llama2-Chat-7B by 6.5% and 6.8% while decreasing those on Zephyr by 4.2% and 2.6%. The performance of DPO compared to GRATH indicates its vulnerability to OOD data.

1. Introduction

With the rapid development of large language models (LLMs), they have been deployed in a wide range of applications (Bommarito II & Katz, 2022; Driess et al., 2023; Lo, 2023). Yet, there is evidence (Bang et al., 2023) showing that LLMs may not answer truthfully, leading to hallucination, *i.e.*, generate factually incorrect content. The hallucination phenomenon could cause great harm, especially in safety-critical applications (Wang et al., 2023; Chen et al., 2023a), underscoring the critical importance of ensuring the models produce truthful outputs. TruthfulQA (Lin et al., 2022) has been considered as a mainstream benchmark for measuring the truthfulness of the model answers to the questions that provoke imitative falsehoods. In particular, MC1 is distinguished as the most challenging task among those built on this benchmark, since even state-of-the-art LLMs could only achieve modest accuracy levels (*e.g.*, the MC1 accuracy of Llama2-Chat-70B (Touvron et al., 2023) is merely 31.09%). These challenges reveal the urgent demand for enhancing the truthfulness of models' outputs.

Existing work (Chuang et al., 2023; Burns et al., 2023) proposes to seek the truthfulness direction within LLMs and shift the model representations along the direction, which leads to truthful outputs. Such process often relies on in-domain data sourced from TruthfulQA (Chen et al., 2023b; Li et al., 2023b). Whereas, in real-world scenarios, it is common to encounter a broad spectrum of data that falls outside the intended testing domain, *i.e.*, out-of-domain (OOD) questions. On the other hand, annotating answers to this extensive range of questions can be prohibitively

[1]UIUC [2]UC Berkeley [3]UChicago. Correspondence to: Weixin Chen <weixinc2@illinois.edu>, Bo Li <bol@uchicago.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

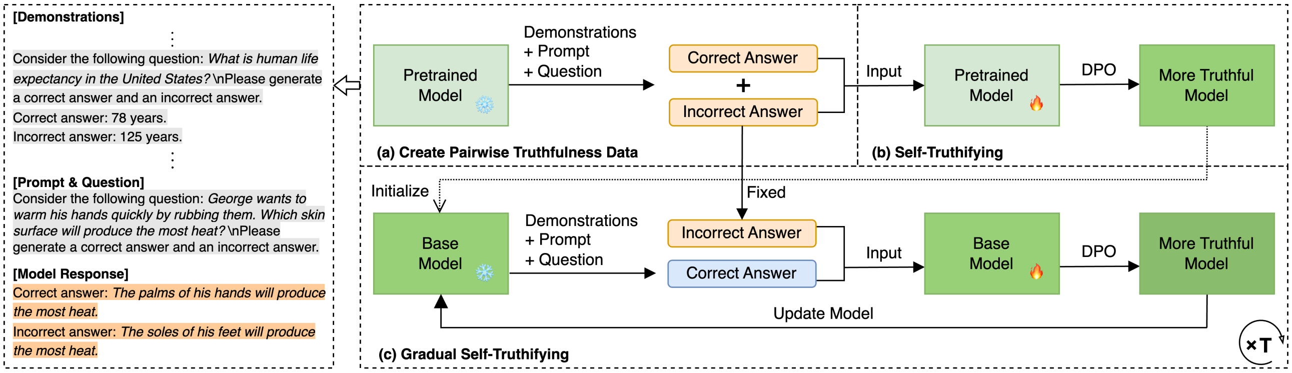

Figure 2. Framework of GRATH, which consists of three components. Given a pretrained base model, GRATH **(a)** creates pairwise truthfulness training data via few-shot prompting. An illustrative example is on the left. A pair of truthfulness data includes a question, a correct answer and an incorrect answer. **(b)** Fine-tune the pretrained model via DPO based on the pairwise truthfulness training data. The model will learn from the truthfulness difference in the self-generated answer pairs and enhance its truthfulness (*i.e.*, self-truthify itself). **(c)** Iteratively generate data and optimize model for T iterations, thus gradually boosting model truthfulness in a self-supervised manner.

expensive and labor-intensive. Furthermore, there exists a risk that these annotations could be noisy or even poisoned (Carlini et al., 2023), leading to a potential decline in model performance. Hence, a natural question arises: *Can we effectively utilize OOD queries to improve the truthfulness of LLMs without needing to rely on human-annotated answers?*

Inspired by the recent alignment technique—Direct Preference Optimization (DPO)—which aligns LLMs with human preferences via learning from the pairwise answers that differ in human alignment level, in this work, we propose a novel post-processing method to boost the truthfulness of pretrained LLMs, named GRAdual self-truTHifying (GRATH), which adaptively generates data using the LLM itself and then optimizes the model via DPO in a self-supervised manner, without the necessity for any annotated answer to the given OOD questions. The pipeline of GRATH is illustrated in Figure 2. First, given a pretrained model, GRATH creates pairwise truthfulness training data for a sequence of questions via few-shot prompting (a pair of truthfulness data is a question and the corresponding correct and incorrect answers). Then, GRATH employs DPO to fine-tune the pretrained model based on the pairwise truthfulness training data, thus enhancing model truthfulness through learning from the difference in answers. Finally, GRATH alternatively refines truthfulness training data and updates the model in an iterative way, leading to the gradual improvement in model truthfulness. As illustrated in Figure 1, when applied to either Llama2-Chat-7B or Zephyr, GRATH could gain a substantial enhancement in both MC1 and MC2 accuracy over the respective pretrained models. In contrast, utilizing OOD questions alongside the annotated pairwise answers, DPO yields a smaller improvement in Llama2-Chat-7B's performance compared to GRATH, and it even results in a deterioration of performance on Zephyr, which reveals its vulnerability to the domain gap.

We conduct intensive experiments, and our empirical results demonstrate that GRATH significantly enhances the MC1 and MC2 accuracy of Llama2-Chat-7B, elevating them from 30.23% to **54.71%** and from 45.32% to **69.10%**, respectively. This marks a new state-of-the-art (SOTA) performance on the TruthfulQA benchmark, surpassing even larger-scale models such as those with 70 billion parameters. Crucially, GRATH achieves this without compromising the pretrained model performance on other established benchmarks, including ARC, HellaSwag, and MMLU. These results underscore GRATH's capability to bolster the truthfulness of LLMs while preserving their core capabilities. Furthermore, we conduct a series of ablation studies and delve into the truthifying processes within GRATH, aiming to gain insights into how to learn more truthfulness.

Our main contributions are threefold. **1)** We discover that the model learned via DPO would be *more* truthful in testing domain if i) the domain gap between pairwise truthfulness training data and testing data is *smaller*; ii) the distributional distance between correct and incorrect answers within pairwise truthfulness training data gets *larger*. **2)** We propose a GRAdual self-truTHifying method, GRATH, to enhance the truthfulness of LLMs in a self-supervised manner. In particular, it adaptively generates pairwise answers to OOD questions and then fine-tunes the model via DPO to improve model truthfulness. **3)** We empirically show that GRATH can significantly improve LLMs' truthfulness, achieving SOTA performance on TruthfulQA's MC1 and MC2 tasks.

2. Related Work

The burgeoning interest in leveraging LLMs to generate text for practical applications necessitates the truthfulness of LLMs. In response to this need, a variety of benchmarks (Li et al., 2023a; Zhang et al., 2023) have been developed to assess the model truthfulness. Notably, TruthfulQA (Lin et al., 2022) has emerged as a prominent benchmark, and its adoption by Open LLM LeaderBoard (Beeching et al., 2023) underscores its significance. In particular, there are multiple-choice questions with a single correct answer in its MC1 task while with a set of correct answers in its MC2 task. A line of research (Chen et al., 2023b; Li et al., 2023b; Chuang et al., 2023; Burns et al., 2023; Zou et al., 2023; Lee et al., 2023a; Azaria & Mitchell, 2023) leverages the models' internal representations (*i.e.,* activations across different layers) to enhance truthfulness. The reliance on representations often necessitates the selection of specific layers. If the optimal layer is not chosen, the enhancement may be compromised. The other line of methods (Joshi et al., 2023) directly focuses on the models' external expressions (*i.e.,* output probabilities) and proposes fine-tuning approaches to improve truthfulness. However, they usually require in-domain data, which is often not accessible in practice. Also, both lines of methods require costly annotated data.

RLHF (Ouyang et al., 2022) and the follow-up work (Bai et al., 2022; Lee et al., 2023b; Gallego, 2023) align LLMs with human preferences by applying reinforcement learning (RL) to optimize the supervised fine-tuned (SFT) model against the reward model trained on pairwise preference data. However, the use of Proximal Policy Optimization (PPO) (Schulman et al., 2017) in RL presents challenges due to the instability and inefficiency. To address this, numerous efficient methods (Dong et al., 2023; Rafailov et al., 2023; Song et al., 2023; Yuan et al., 2023) have been proposed. In particular, DPO develops an objective function to directly optimize the model to adhere to pairwise preference data, avoiding the need to fit a reward model. It is shown better than RLHF while easier to implement and needs fewer resources, making it the focal point of our paper. Nevertheless, the domain gap issue inherent in DPO (*i.e.*, the vulnerability to domain shifts) has not yet been thoroughly studied, which could potentially lead to performance degradation. On the other hand, it still requires a large amount of annotations for pairwise preference data, which could be costly. In practical scenarios, we might only have limited annotated data.

Concurrently, there are several works also leveraging self-generated data to improve LLMs' capabilities. Different from them, we focus on enhancing model truthfulness. Unlike SPIN (Chen et al., 2024), we do not require a large human-annotated dataset to initialize an SFT model, nor do we rely on an external scalar feedback signal as a quality indicator like ReST^{EM} (Singh et al., 2023) or utilize LLM to reward answers as in Self-Rewarding (Yuan et al., 2024), since our primary concern is the relative truthfulness between generated correct and incorrect answers, rather than the absolute truthfulness of each. As opposed to GAN-like training in SPIN or SFT in ReST^{EM}, we utilize DPO, with a specific focus on exploring the domain gap issue in DPO.

3. Method

We propose a GRAdual self-truTHifying method for LLMs, named GRATH, which gradually enhances the model truthfulness in a self-supervised manner. As illustrated in Figure 2, GRATH consists of three components, 1) *creating pairwise truthfulness data*, 2) *self-truthifying*, and 3) *gradual self-truthifying*. In this section, we will elaborate on the mechanisms of different components, demonstrating how to obtain a truthful model given a pretrained base model.

3.1. Creating Pairwise Truthfulness Data

Annotating correct and incorrect answers for large-scale questions could take a substantial amount of human effort and resources. Hence, we attempt to prompt the model to generate a correct answer a_T and an incorrect answer a_F to any given question q.

Questions. Assume that the ultimate goal for the model is to gain high truthfulness on questions from a target domain $\mathcal{D}_{\text{target}}$. In practice, we hardly have access to $\mathcal{D}_{\text{target}}$ since we do not know which domain the model would be tested on. Therefore, we randomly select an open-source dataset consisting of questions, *i.e.*, $\{q^i\}_{i=1}^n \subseteq \mathcal{D}_{\text{source}}$. Note that in practice, the dataset is usually out-of-domain in terms of the target domain.

Prompt. We design the following prompt $p(\cdot)$ intriguing the model to generate a correct answer a_T^i and an incorrect answer a_F^i to any question q^i, constituting a pair of truthfulness training data (q^i, a_T^i, a_F^i).

> $p(q^i)$ = "Consider the following question: q^i\nPlease generate a correct answer and an incorrect answer. Make sure the answers are plausible. There is no need to give an explanation."

Few-shot Demonstrations. To ensure the model generates responses in the desired format, we include a few demonstrations $\{(\hat{q}^j, \hat{a}_T^j, \hat{a}_F^j)\}_{j=1}^m$ before the above prompt. An illustrative template $\mathbb{T}(p(\cdot); \{(\hat{q}^j, \hat{a}_T^j, \hat{a}_F^j)\}_{j=1}^m)$ along with the model response is shown in Figure 2. Detailed templates are illustrated in Figure 10, 11, 12 in Appendix B. We also give some examples of the created pairwise truthfulness training data in Figure 13 in Appendix C.

3.2. Self-Truthifying

Recent offline alignment methods (Rafailov et al., 2023; Song et al., 2023) align LLMs with humans by increasing the probability of generating responses annotated with higher quality while decreasing that annotated with lower quality. That is, the alignment of the model is improved via learning from the responses that differ in alignment quality.

Inspired by this, we attempt to learn from the responses that differ in truthfulness, thus enhancing the truthfulness of the model. Specifically, we leverage an effective offline alignment method DPO (Rafailov et al., 2023) to fine-tune the pretrained base model using the created pairwise truthfulness training data.

Formally, let π_θ denote the model to be fine-tuned where θ are learnable parameters, π_{ref} denote the fixed reference model, and $D_{\text{pair}} := \{(q^i, a^i_T, a^i_F)\}_{i=1}^n$ denote all pairwise truthfulness training data. The loss function of DPO, *i.e.*, $\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}, D_{\text{pair}})$, is given by:

$$-\frac{1}{n}\sum_{i=1}^n \left[\log\sigma\left(\beta\log\frac{\pi_\theta\left(a^i_T|q^i\right)}{\pi_{\text{ref}}\left(a^i_T|q^i\right)} - \beta\log\frac{\pi_\theta\left(a^i_F|q^i\right)}{\pi_{\text{ref}}\left(a^i_F|q^i\right)}\right)\right],$$

where σ is the logistic function and β serves as a parameter that regulates the deviation from the reference model π_{ref}, with both π_θ and π_{ref} being initialized as the pretrained base model π_{pre}.

By minimizing the loss function, the likelihood of generating correct (incorrect) answers will increase (decrease), thus contributing to a more truthful model. Since the model enhances its truthfulness by learning from the difference in answer pairs that are generated by itself without requiring any external annotations, we name the learning process as *self-truthifying*.

3.3. Gradual Self-Truthifying

Gradual self-truthifying consists of two alternative steps—refining the data and updating the model. These steps are executed alternatively in an iterative manner, contributing to the gradual improvement in model truthfulness.

Step 1: Refining Data. Intuitively, the model with improved truthfulness could generate correct answers with higher correctness, which inherently benefits the process of learning from the truthfulness difference in answer pairs. Hence, regarding the self-truthified model as the base model, we prompt it to generate correct answers and substitute those in the pairwise truthfulness training data while leaving the incorrect answers fixed.

Step 2: Updating Model. Based on the refined pairwise truthfulness training data, we employ self-truthifying on the base model (*i.e.*, fine-tune the base model via DPO), aiming to improve its truthfulness by learning from the truthfulness difference in the answer pairs. Subsequently, the learned model will be used to update the base model.

The above process of refining the data and updating the model could be repeated iteratively until a predetermined number of iterations is reached. Through this iterative approach, we could gradually boost the model truthfulness in a self-supervised manner. The overall procedure of GRATH is summarized in Algorithm 1.

Algorithm 1 GRATH

Input: a pretrained model π_{pre}, a set of questions $\{q^i\}_{i=1}^n$, few-shot demonstrations $\{(\hat{q}^j, \hat{a}^j_T, \hat{a}^j_F)\}_{j=1}^m$, prompting template $\mathbb{T}(p(\cdot);\cdot)$, number of fine-tuning steps s, number of iterations T

Output: a model with high truthfulness π^*_θ

 # (a) Create pairwise truthfulness data

$(a^i_T, a^i_F) = \pi_{\text{pre}}\left(\mathbb{T}(p(q^i); \{(\hat{q}^j, \hat{a}^j_T, \hat{a}^j_F)\}_{j=1}^m)\right)$ for i in $[n]$

$D^0_{\text{pair}} = \{(q^i, a^i_T, a^i_F)\}_{i=1}^n$

 # (b) Self-truthifying

$\pi^0_\theta \leftarrow \pi_{\text{pre}}$, $\pi_{\text{ref}} \leftarrow \pi_{\text{pre}}$

$\pi^1_\theta \leftarrow \pi^0_\theta$ fine-tuned with $\mathcal{L}_{\text{DPO}}\left(\pi^0_\theta; \pi_{\text{ref}}, D^0_{\text{pair}}\right)$ for s steps

 # (c) Gradual self-truthifying

for t in $[T]$ **do**

  $(\tilde{a}^i_T, \tilde{a}^i_F) = \pi^t_\theta\left(\mathbb{T}(p(q^i); \{(\hat{q}^j, \hat{a}^j_T, \hat{a}^j_F)\}_{j=1}^m)\right)$ for i in $[n]$

  $D^t_{\text{pair}} = \{q^i, \tilde{a}^i_T, a^i_F\}_{i=1}^n$

  $\pi_{\text{ref}} \leftarrow \pi^t_\theta$

  $\pi^{t+1}_\theta \leftarrow \pi^t_\theta$ tuned with $\mathcal{L}_{\text{DPO}}\left(\pi^t_\theta; \pi_{\text{ref}}, D^t_{\text{pair}}\right)$ for s steps

end for

return $\pi^*_\theta = \pi^{T+1}_\theta$

4. Experiments

4.1. Experimental Setup

Models. We adopt two widely-used 7B-LLMs as pretrained base models, which are Llama2-Chat-7B (Touvron et al., 2023) and Zephyr (Tunstall et al., 2023). Each is used to create pairwise truthfulness training data and initialize the base model in self-truthifying. During gradual self-truthifying, this self-truthified model is used as the base model in the first iteration, and subsequently, the model trained by DPO will serve as the base model in the next iteration. Following (Tunstall et al., 2023), we compare GRATH against a variety of open-access models featured on Open LLM Leaderboard (Beeching et al., 2023) with scales of parameters ranging from 7B to 70B, including Xwin-LM (Ni et al., 2024), Mistral-Instruct (Jiang et al., 2023), MPT-Chat (Team, 2023), StableLM-α, Llama2-Chat (Touvron et al., 2023), Zephyr (Tunstall et al., 2023), Vicuna (Chiang et al., 2023), WizardLM (Xu et al., 2023), and Guanaco (Dettmers et al., 2023).

Baseline Methods. In addition to the above open models, we also compare GRATH with various fine-tuning methods which cover two main categories: *human alignment* methods—SFT, DPO (Rafailov et al., 2023), RLHF (Ouyang et al., 2022) and *truthfulness enhancement* methods—RepE (Zou et al., 2023), ITI (Li et al., 2023b). To apply human alignment methods within the realm of truthfulness, we implement them using the same dataset as ours, *i.e.,* ARC-

Challenge. Different from ours, they would utilize both the questions and the annotated answers. Besides, we report ITI's results using its published model and reproduce RepE using its open-source codes. All of the methods use LLama2-Chat-7B as the pretrained base model.

Datasets and Evaluation Metrics. When creating pairwise truthfulness training data, we utilize training samples from ARC-Challenge (Clark et al., 2018) as the source of questions, while using six QA primer examples from TruthfulQA as few-shot demonstrations which follows the implementation in (Zou et al., 2023). The core capabilities of the models are assessed using testing samples from ARC-Challenge, HellaSwag (Zellers et al., 2019), and MMLU (Hendrycks et al., 2021). Meanwhile, the truthfulness of the models is evaluated on TruthfulQA's MC tasks (Lin et al., 2022), where we utilize the entire dataset for assessment unless otherwise specified. We follow the evaluation configurations and procedures introduced in Open LLM Leaderboard, adopting accuracy as the metric. Specifically, MC1 accuracy is the percentage of assigning the highest probability to the single correct answer in TruthfulQA's MC1 task. MC2 accuracy is the normalized total probability assigned to a set of correct answers in the MC2 task.

Implementation Details. We prompt the pretrained model to generate pairwise answers to all questions in ARC-C. The model generations that do not follow the format defined in demonstrations are filtered out, which results in 1092 pairs of truthfulness training data in total. We adopt DPO to fine-tune the model for 1000 steps using the parameter-efficient technique—LoRA (Hu et al., 2022). Empirically, we find out that a single iteration $T = 1$ suffices to yield substantial improvements in truthfulness, signifying GRATH as an efficient and convenient post-processing approach for enhancing LLMs' truthfulness. More details about the experimental setup are presented in Appendix A.

4.2. Effectiveness of GRATH

4.2.1. Comparison with Open Models

The performance of a variety of open-access models and GRATH over two pretrained base models on different benchmark datasets are shown in Table 1.

Compared with the corresponding pretrained base models, GRATH marginally improves performance on ARC-C, albeit with a slight trade-off in performance on HellaSwag and MMLU. These results suggest that GRATH could preserve the performance on these datasets, maintaining the models' core capabilities.

On the TruthfulQA benchmark, we observe that GRATH effectively improves the MC1 and MC2 accuracy, with both increased by an average of over 15%. In particular, GRATH$_{\text{Llama2}}$ achieves the state-of-the-art (SOTA) performance, with impressive MC1 accuracy of 54.71% and MC2 accuracy of 69.10%. Notably, this 7B model's performance even outstrips that of larger-scale models, including those with 70B parameters. For instance, its MC1 and MC2 accuracy exceed those of the leading 70B model (*i.e.*, Xwin-LM v0.1) by substantial margins of 14.44% and 9.44%, respectively. Meanwhile, GRATH$_{\text{Zephyr}}$ also demonstrates superior performance compared to the selected open models. More experimental results are shown in Appendix D.1. These results indicate that **GRATH serves as an effective post-processing method to bolster the truthfulness of different LLMs with minimal impact on their core capabilities.**

Table 1. Performance of GRATH compared with various open models on Open LLM Leaderboard. On TruthfulQA's MC1 and MC2 tasks, GRATH$_{\text{Llama2}}$ achieves **SOTA** performance, and GRATH$_{\text{Zephyr}}$ ranks as the <u>second best</u>. On other core functionality benchmarks, GRATH maintains similar performance compared to the corresponding pretrained base models.

Model	Size	ARC	Hella Swag	MMLU	TQA MC1	TQA MC2
StableLM-Tuned-α	7B	31.91	53.59	24.41	23.99	40.37
MPT-Chat	7B	46.50	75.51	37.62	27.05	40.16
Xwin-LM v0.1	7B	56.57	79.40	49.98	32.93	47.89
Mistral-Instruct v0.1	7B	54.52	75.63	55.38	39.53	56.28
Vicuna v1.3	33B	62.12	83.00	59.22	37.09	56.16
Guanaco	65B	65.44	86.47	62.92	36.47	52.81
Llama2-Chat	70B	<u>67.32</u>	**87.33**	**69.83**	31.09	44.92
WizardLM v1.0	70B	65.44	84.41	64.05	38.68	54.81
Xwin-LM v0.1	70B	**70.22**	<u>87.25</u>	<u>69.77</u>	40.27	59.86
Zephyr	7B	62.46	84.35	60.70	42.23	57.83
GRATH$_{\text{Zephyr}}$	7B	65.02	81.57	51.39	<u>53.86</u>	<u>66.73</u>
Llama2-Chat	7B	52.73	78.50	48.14	30.23	45.32
GRATH$_{\text{Llama2}}$	7B	57.76	79.63	46.88	**54.71**	**69.10**

4.2.2. Comparison with Baseline Methods

The performance comparison between different fine-tuning methods is illustrated in Table 2. These methods include human alignment methods (SFT, DPO, RLHF) and truthfulness enhancement methods (RepE, ITI, GRATH). Note that RLHF employs RL to fine-tune the SFT model obtained in its first stage. To study how this fine-tuning paradigm learns truthfulness, we also apply RL on pretrained base model, excluding the influence of SFT base model, denoted as RL.

Among various human alignment methods, DPO performs the best, as it not only slightly increases the average accuracy on ARC-C, HellaSwag, and MMLU datasets, but also boosts the average accuracy on TruthfulQA's MC tasks by 6.65%. In comparison, RL barely affects the pretrained model's performance while both SFT and RLHF tend to diminish the performance, which indicates that SFT might not be an appropriate strategy for fine-tuning models with OOD training data. When it comes to methods aimed at enhancing truthfulness, RepE is superior to DPO, effectively increasing the average accuracy on TruthfulQA by 14.18%. Moreover, GRATH even outperforms RepE, delivering a substantial improvement of 24.14% while maintaining performance on the other three datasets. These results again validate GRATH's effectiveness in enhancing the model truthfulness.

Table 2. Performance comparison between human alignment methods—SFT, DPO, RMRL, RL, and truthfulness enhancement methods—RepE, ITI, GRATH. All methods are applied on Llama2-Chat-7B. On TruthfulQA's MC1 and MC2 tasks, $\text{GRATH}_{\text{Llama2}}$ achieves **SOTA** performance, and RepE ranks as the second best. On other core functionality benchmarks, most methods maintain similar performance compared to Llama2-Chat-7B.

Method	ARC	Hella Swag	MMLU	AVG	TQA MC1	TQA MC2	AVG
$\text{Llama2-Chat}_{7\text{B}}$	52.73	78.50	48.14	59.79	30.23	45.32	37.77
SFT	51.88	76.65	45.89	58.14	22.40	34.52	28.46
DPO	**62.97**	**81.84**	44.01	**62.94**	36.72	52.13	44.42
RL	53.41	78.36	**48.30**	60.03	30.11	44.89	37.50
RLHF	49.91	76.56	45.14	57.20	22.64	34.37	28.50
RepE	57.59	78.75	47.86	61.40	43.45	60.45	51.95
ITI	52.73	78.50	48.17	59.80	30.23	45.32	37.77
$\text{GRATH}_{\text{Llama2}}$	57.76	79.63	46.88	61.42	**54.71**	**69.10**	**61.91**

4.3. In-Depth Exploration of GRATH

In this section, we will delve into GRATH to comprehend how it facilitates the learning of a truthful model. We first introduce different disparity concepts that will be instrumental in our analysis.

Definition 1 (Domain Gap). The domain gap refers to the disparity between the training domain and the testing domain. More precisely, it refers to the disparity between the pairwise truthfulness training data and the testing data.

Definition 2 (Distributional Distance). The distributional distance within pairwise truthfulness training data refers to the disparity between the distribution of correct answers and that of incorrect answers.

Definition 3 (Pairwise Distance). The pairwise distance between a correct answer a_T and an incorrect answer a_F is the Euclidean distance between their vectors projected onto a representation space $\mathcal{M}$, *i.e.*, $d_{\text{pair}}(a_T, a_F) = \|\mathcal{M}(a_T) - \mathcal{M}(a_F)\|_2$.

In particular, we use the representation at the last token in the last layer of the Llama2-Chat-7B model. Consequently, the distributional distance can be characterized by the statistics of $\{d_{\text{pair}}(a_T^i, a_F^i)\}_{i=1}^n$ (*e.g.*, mean).

4.3.1. Towards Understanding How to Learn Truthfulness in Self-Truthifying

In this section, we aim to investigate the factors that affect the truthfulness of the model learned in *self-truthifying*.

— *Does the domain gap between pairwise truthfulness training data and testing data impact the truthfulness learned by DPO?* During evaluation, the truthfulness of the learned model is assessed in a particular testing domain. Herein, we explore how this truthfulness is influenced when the domain gap in the training data is widened or narrowed.

The model learned by DPO is more truthful in the testing domain if there is a smaller domain gap between pairwise truthfulness training data and testing data. First, we split TruthfulQA into 700 training samples and 117 testing samples. One training sample (*i.e.*, a pair of truthfulness data) comprises a question, a correct answer, and a randomly selected incorrect answer. To elevate the degree of domain gap with respective to the testing domain, we employ a sentence-level transformation (Krishna et al., 2020) on the answers in training samples. In particular, we transform the answers to the Shakespearean style with different levels of perturbations characterized by the parameter top-p. A larger top-p indicates higher level of perturbation, thereby increasing the domain gap in the answers. Note that top-p = 0.0 equates to no transformation, hence no domain gap. As illustrated in Figure 3(left), there is a clear downward trend in both MC1 and MC2 accuracy as the domain gap in answers widens. This pattern demonstrates that 1) when there is no domain gap in questions, the model learned by DPO will be less truthful if the domain gap in answers increases.

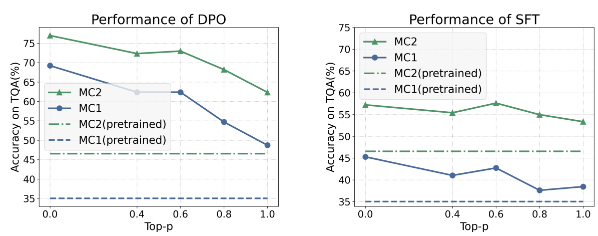

Figure 3. MC1 and MC2 accuracy of DPO (left) and SFT (right) with varying degrees of transformations applied on the answers in pairwise truthfulness training data. A larger top-p indicates a larger domain gap between truthfulness training data and testing data. Downward trends here in each figure indicate that the model learned by either DPO or SFT will be less truthful if the domain gap in answers increases. DPO performs better than SFT overall.

In Section 4.2, DPO is applied on the pairwise truthfulness data created using OOD questions (from ARC-C) and 6 in-domain few-shot demonstrations (from TruthfulQA). We denote such DPO which leverages the model's generated answers as $DPO^{Q_{OOD}}_{G(FS_{IND})}$. Here, we extend this creating procedure to more diverse setups, including: i) merging the ground-truth correct and incorrect answers annotated in ARC-C into the prompt during model generation (denoted as $DPO^{Q_{OOD}}_{G(FS_{IND},GT)}$); ii) altering the few-shot demonstrations to OOD examples sourced from HaluEval dataset (Li et al., 2023a) (denoted as $DPO^{Q_{OOD}}_{G(FS_{OOD})}$); iii) combining

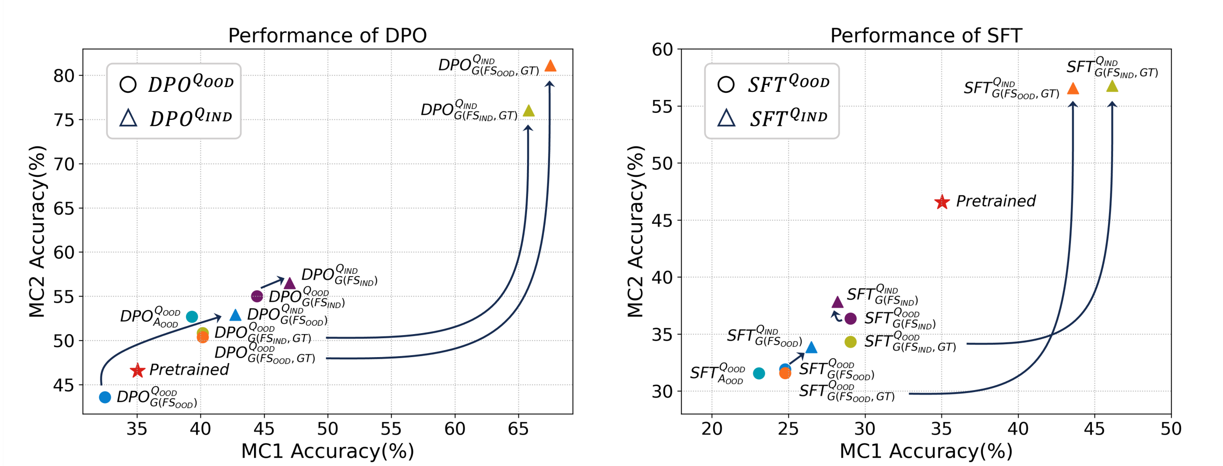

Figure 4. MC1 and MC2 accuracy of DPO (left) and SFT (right). DPO and SFT are applied with truthfulness training data created using a variety of strategies. Q_{OOD} (Q_{IND}) represents using OOD (in-domain) questions from ARC-C (TruthfulQA); A_{OOD} indicates using annotated answers from ARC-C; $G(\cdot)$ indicates using answers generated by LLMs. Specifically, FS_{OOD} (FS_{IND}) corresponds to OOD (in-domain) few-shot demonstrations, and GT implies merging ground-truth answers into the prompts. We find i) $DPO^{Q_{OOD}}_{G(FS_{IND})}$ performs the best among all $DPO^{Q_{OOD}}$, indicating that the usage of in-domain demonstrations yields answers that are closer to testing domain, leading to a more truthful model. ii) Arrows symbolize performance shifts as questions are transitioned from OOD to in-domain. The trends towards upper right indicate that the model will be more truthful if the domain gap in questions decreases. The same findings hold for SFT. iii) DPO outperforms SFT since it improves the pretrained model's truthfulness in general.

both of them (denoted as $DPO^{Q_{OOD}}_{G(FS_{OOD},GT)}$). In addition, we consider directly feeding the annotated pairwise answers in ARC-C to DPO which is denoted as $DPO^{Q_{OOD}}_{A_{OOD}}$. For a fair comparison, we uniformly utilize 700 truthfulness training data across all strategies. Note that without the usage of demonstrations, the model might fail to generate 700 truthfulness data. Thus, we do not study this zero-shot setting. As depicted by the circles in Figure 4(left), we discover that $DPO^{Q_{OOD}}_{G(FS_{IND})}$ (shown by the purple circle) performs the best among all $DPO^{Q_{OOD}}$. We infer that its utilization of in-domain demonstrations potentially yields answers that are closer to the testing domain. In contrast, other strategies that either refer to the OOD ground-truth answers or switch to OOD demonstrations might cause the generated answers to deviate from the testing domain, thus impairing performance in comparison to $DPO^{Q_{OOD}}_{G(FS_{IND})}$. The results indicate that 2) when there is a domain gap in questions, the model learned by DPO will be less truthful if the domain gap in answers increases.

So far, we have discussed the impact of the domain gap in answers on model truthfulness. To further explore the effects of the domain gap in questions, we implement the same four answer generation strategies, but using in-domain questions (from TruthfulQA). Arrows in Figure 4(left) symbolize the performance shifts as the questions are transitioned from OOD to in-domain under each respective strategy. Notably, there is a trend toward the upper right, signifying a general improvement when the questions become in-domain, which demonstrates that 3) the model learned by DPO will be more truthful if the domain gap in questions decreases.

— *Can SFT be a better choice than DPO in terms of learning truthfulness?* In addition to studying the effects of training data, here we focus on the fine-tuning technique. We are intrigued by whether the widely-used approach—Supervised Fine-Tuning (SFT)—offers comparable benefits with DPO.

The model learned via SFT is less truthful in the testing domain if there is a larger domain gap between pairwise truthfulness training data and testing data. Besides, SFT is less effective than DPO for improving truthfulness. Here, we replicate the setups previously used, with the difference that SFT only utilizes correct answers instead of pairwise answers. Specifically, we first apply varying levels of transformations on the annotated correct answers from TruthfulQA. The downward trends of accuracy in Figure 3(right) demonstrate that 1) when there is no domain gap in questions, the model learned by SFT will be less truthful if the domain gap in answers increases. Next, we follow the same answer generation strategies given questions from ARC-C, while applying the generated correct answers to SFT. As shown in Figure 4(right), $SFT^{Q_{OOD}}_{G(FS_{IND})}$ (shown by the purple circle) outperforms other $SFT^{Q_{OOD}}$ (represented by circles in other colors), suggesting that 2) when there is a domain gap in questions, the model learned by SFT will be less truthful if the domain gap in answers increases. Then, upon switching the questions from OOD to in-domain, the trend of most arrows toward the upper right signifies that 3) the model learned by SFT will be more truthful if the domain gap in questions decreases. Notably, we discover that the majority of DPO's points are situated to the upper right of the pretrained model (shown by the red star), signifying enhanced model truthfulness. By contrast, most SFT's points are located in the lower left of the pretrained model, suggesting degraded model truthfulness. The comparison indicates that 4) DPO is more effective than SFT for improving model truthfulness.

— *How does the number of fine-tuning steps affect the model truthfulness as well as other core capabilities?* We defer the discussion to Appendix D.2.

4.3.2. Towards Understanding How to Learn Truthfulness in GRATH

In this section, we aim to investigate the factors that affect the truthfulness learned in *GRATH*. Following the notations in Section 3, T denotes the iterations in gradual self-truthifying. To avoid ambiguity, let T_{DPO} denote the times that DPO is executed. Apparently, $T_{DPO} = T + 1$. In particular, DPO^0 refers to the scenario where no DPO is performed, meaning the model remains in its pretrained state. DPO^1 represents the DPO in self-truthifying while DPO^t $(1 < t \leq T_{DPO})$ corresponds to its use in gradual self-truthifying.

— *Does the distributional distance between correct and incorrect answers within pairwise truthfulness training data impact the truthfulness learned by* ***DPO****?* In GRATH, we fine-tune the model using DPO two times since $T = 1$. Then, a natural question arises: Compared to DPO^1, why the model can be more truthful after DPO^2? Is it only because the base model is more truthful? Or is it also related to the quality of truthfulness data used in DPO^2?

The model learned via DPO is more truthful if the distributional distance between correct and incorrect answers within pairwise truthfulness training data is larger. Figure 5(left) depicts the distributions of pairwise distance in truthfulness data used in DPO^1 and DPO^2. There is a distinct shift towards right when transitioning from DPO^1 to DPO^2, with the mean value evolving from 65.94% to 87.80%. This trend indicates that the truthfulness data in DPO^2 exhibits a larger distributional distance. To further validate whether truthfulness data with a larger distributional distance contributes to a more truthful model, we fine-tune the pretrained model with DPO using these two datasets, respectively. As illustrated in Figure 5(right), both MC1 and MC2 accuracy present an uplift, verifying that a larger distributional distance within truthfulness data leads to a higher degree of truthfulness in the model learned via DPO.

— *Can the alignment approach adopted in GRATH be extended from DPO to various direct alignment from preference (DAP) approaches?* DAP approaches aim to align a policy directly with preference pairs, with DPO being a representative one. To explore the effectiveness of GRATH applied with different DAP approaches, we replace DPO with two representative DAP approaches: IPO (Azar et al., 2024) and RSO (Liu et al., 2024), respectively, and demonstrate the performance of the resulting models in Table 3. In particular, the models obtained after self-truthifying and an iteration of gradual self-truthifying are denoted by $\text{GRATH}^{\text{DAP}^1}_{\text{Llama2}}$ and $\text{GRATH}^{\text{DAP}^2}_{\text{Llama2}}$, respectively.

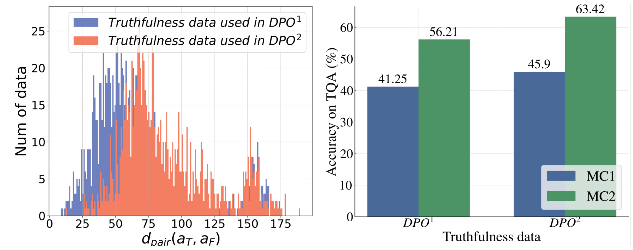

Figure 5. Left: Distributions of pairwise distance in truthfulness data used in $DPO^{1,2}$. Right: Performance of the pretrained model fine-tuned with truthfulness data used in $DPO^{1,2}$ on TruthfulQA. The improved performance indicates that a larger distributional distance within truthfulness data leads to a more truthful model.

The alignment approach adopted in GRATH can be generalized from DPO to various direct alignment from preference approaches. Compared to the pretrained base model, all of the DAP approaches, *i.e.*, DPO, IPO, and RSO, could effectively increase the average accuracy on the TruthfulQA benchmark by over 9% in self-truthifying. During gradual self-truthifying, they could further improve the truthfulness of the corresponding self-truthified models, leading the average accuracy on the TruthfulQA benchmark to 61.91%, 53.51%, and 49.13%, respectively. Meanwhile, their performance on other core functionality benchmarks maintains similar to that of the pretrained base model. Therefore, GRATH is generalizable over different DAP approaches.

Table 3. Performance comparison between the pretrained base model (*i.e.*, Llama2-Chat-7B) and GRATH using various direct alignment from preference (DAP) approaches. $\text{GRATH}^{\text{DAP}^1}_{\text{Llama2}}$ and $\text{GRATH}^{\text{DAP}^2}_{\text{Llama2}}$ denote the models obtained after self-truthifying and an iteration of gradual self-truthifying, respectively.

Method	ARC	Hella Swag	MMLU	AVG	TQA MC1	TQA MC2	AVG
Llama2-Chat$_{7B}$	52.73	78.50	48.14	59.79	30.23	45.32	37.77
$\text{GRATH}^{\text{DPO}^1}_{\text{Llama2}}$	56.40	79.68	47.22	61.10	41.25	56.21	48.73
$\text{GRATH}^{\text{DPO}^2}_{\text{Llama2}}$	57.76	79.63	46.88	61.42	54.71	69.10	61.91
$\text{GRATH}^{\text{IPO}^1}_{\text{Llama2}}$	57.85	78.49	44.34	60.23	41.49	60.77	51.13
$\text{GRATH}^{\text{IPO}^2}_{\text{Llama2}}$	57.34	79.10	44.46	60.30	44.43	62.58	53.51
$\text{GRATH}^{\text{RSO}^1}_{\text{Llama2}}$	54.61	79.19	47.83	60.54	38.43	56.97	47.70
$\text{GRATH}^{\text{RSO}^2}_{\text{Llama2}}$	56.74	79.50	46.33	60.86	39.90	58.37	49.13

— *What is the relationship between the number of DPO executions T_{DPO} and the amount of pairwise truthfulness training data n?* To answer the question, we fix the number of fine-tuning steps as 1000 in each DPO execution and conduct GRATH with varying T_{DPO} and n.

GRATH is an efficient approach to boost model truthfulness. In particular, a larger amount of pairwise truthfulness training data leads to a more truthful model. As shown in Figure 6, under a fixed T_{DPO}, GRATH with fewer samples tends to exhibit lower performance, as reflected in the diminished MC1 and MC2 accuracy. Note that MC2 accuracy could be NaN, which is possibly due to that the numerous executions lead the model over-fitting to pairwise truthfulness data, at the expense of deviating from the pretrained model. This deviation might hinder the model from generating fluent text, resulting in an extremely low probability of generating fluent text. Consequently, MC2 accuracy, defined as the normalized probability assigned to a set of correct answers, might be NaN. And for a fixed amount of truthfulness data n, GRATH typically achieves its peak MC1 accuracy when $T_{DPO} = 3$ or 4. Notably, this optimal accuracy declines as n reduces. These results demonstrate that GRATH only requires $T_{DPO} \leq 4$ to attain its highest level of MC1 accuracy, with larger amounts of truthfulness data yielding higher degree of truthfulness.

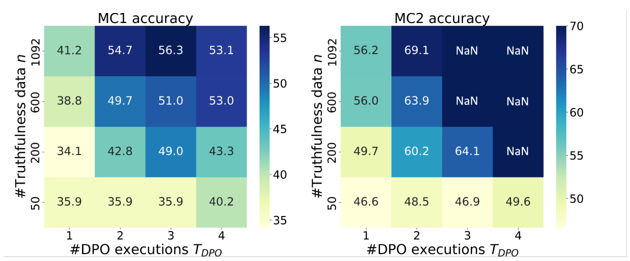

Figure 6. MC1 (left) and MC2 (right) accuracy of GRATH with different amounts of truthfulness data n and numbers of DPO executions T_{DPO}. For a fixed T_{DPO}, a larger n leads to a more truthful model. MC2 accuracy might be NaN due to over-fitting.

— *How does the number of iterations affect the model truthfulness as well as other core capabilities? What is the impact of the choice of DPO's reference model?* We defer the discussion to Appendix D.3.

5. Discussion

Overfitting to Pairwise Data. As the training progresses, the learned model might overfit to the pairwise training data (Azar et al., 2024). This overfitting could affect some capabilities implicitly learned during pretraining, as implied by the performance drop on core functionality benchmarks in later iterations during gradual self-truthifying, which is shown in Figure 8 in Appendix D.3. This impact might lead to generations of lower quality in certain aspects, potentially affecting the truthfulness of the fine-tuned model.

Here, we provide some potential solutions to overfitting. 1) Early stopping: Such strategies include reducing the steps in DPO or the number of truthifying iterations. For instance, as shown in Figure 8, there is minimal impact on the performance on core functionality benchmarks when the model is truthified for one or two iterations. 2) Setting DPO's reference model as the pretrained base model: As specified in Section 3, DPO's reference model is set as the current base model in each truthifying iteration, which may accelerate model's overfitting to pairwise data. In Appendix D.3, we explore *how the choice of reference model in DPO affects the performance of the learned model*, where we fix DPO's reference model as the pretrained base model. Results demonstrate that adding regularization towards the pretrained base model can facilitate stable performance on core functionality benchmarks throughout the training.

Enhancement of LLMs' Multiple Capabilities. While the focus of this paper is on truthfulness, LLMs are expected to exhibit a range of desirable qualities. A natural question arises: *How can we enhance LLMs' multiple capabilities simultaneously?* A naive idea would be to extend the creation procedure in GRATH to generate pairwise answers that differ in diverse attributes, *e.g.*, harmlessness and morality.

Due to space limit, we defer the discussion on the evaluation of model truthfulness to Appendix E.

6. Conclusion

In this paper, we propose GRATH, a novel post-processing method to enhance LLMs' truthfulness. Intensive experiments have demonstrated GRATH's *effectiveness*, *OOD-resilience*, *cost-effectiveness*, and *efficiency*. Furthermore, beyond the methodology itself, we provide an innovative fine-tuning paradigm, *i.e.*, gradual self-training, which leverages the self-generated data to improve a model's certain ability, while utilizing the model's improved ability to generate data of better quality. The alternative interaction between them contributes to the gradual enhancement in the ability. Diverse alignment approaches can be incorporated into this paradigm, as evidenced by our exploration of DPO, SFT, IPO, and RSO. Such *flexibility* enables it to capitalize on the distinct benefits of different approaches, potentially offering widespread implications for LLMs.

Acknowledgements. This work is partially supported by the National Science Foundation under grant No. 1910100, No. 2046726, No. 2229876, DARPA GARD, the National Aeronautics and Space Administration (NASA) under grant no. 80NSSC20M0229, the Alfred P. Sloan Fellowship, and the Amazon research award.

Impact Statement

Our work aims to enhance the truthfulness of LLMs, and we focus on reducing the possibility of LLMs generating misinformation. We expect our work will not have negative societal consequences if there is any. Overall, fully equipping LLMs with comprehensive trustworthiness attributes is critical before large-scale LLM deployment in the real world, which still remains an unresolved challenge, and our work aims to push the frontier of it. Addressing this challenge is crucial for further enhancing the reliability of LLMs across various applications.

References

Azar, M. G., Guo, Z. D., Piot, B., Munos, R., Rowland, M., Valko, M., and Calandriello, D. A general theoretical paradigm to understand learning from human preferences. In AISTATS, 2024.

Azaria, A. and Mitchell, T. M. The internal state of an LLM knows when it's lying. In EMNLP, 2023.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. Constitutional ai: Harmlessness from ai feedback. arXiv preprint arXiv:2212.08073, 2022.

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. arXiv preprint arXiv:2302.04023, 2023.

Beeching, E., Fourrier, C., Habib, N., Han, S., Lambert, N., Rajani, N., Sanseviero, O., Tunstall, L., and Wolf, T. Open llm leaderboard. https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard, 2023.

Bommarito II, M. and Katz, D. M. Gpt takes the bar exam. arXiv preprint arXiv:2212.14402, 2022.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. In NeurIPS, 2020.

Burns, C., Ye, H., Klein, D., and Steinhardt, J. Discovering latent knowledge in language models without supervision. In ICLR, 2023.

Carlini, N., Jagielski, M., Choquette-Choo, C. A., Paleka, D., Pearce, W., Anderson, H., Terzis, A., Thomas, K., and Tramèr, F. Poisoning web-scale training datasets is practical. arXiv preprint arXiv:2302.10149, 2023.

Chen, L., Sinavski, O., Hünermann, J., Karnsund, A., Willmott, A. J., Birch, D., Maund, D., and Shotton, J. Driving with llms: Fusing object-level vector modality for explainable autonomous driving. arXiv preprint arXiv:2310.01957, 2023a.

Chen, Z., Sun, X., Jiao, X., Lian, F., Kang, Z., Wang, D., and Xu, C.-Z. Truth forest: Toward multi-scale truthfulness in large language models through intervention without tuning. arXiv preprint arXiv:2312.17484, 2023b.

Chen, Z., Deng, Y., Yuan, H., Ji, K., and Gu, Q. Self-play fine-tuning converts weak language models to strong language models. arXiv preprint arXiv:2401.01335, 2024.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., and Xing, E. P. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

Chuang, Y.-S., Xie, Y., Luo, H., Kim, Y., Glass, J., and He, P. Dola: Decoding by contrasting layers improves factuality in large language models. arXiv preprint arXiv:2309.03883, 2023.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. Think you have solved question answering? try arc, the ai2 reasoning challenge. arXiv preprint arXiv:1803.05457, 2018.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. Qlora: Efficient finetuning of quantized llms. arXiv preprint arXiv:2305.14314, 2023.

Dong, H., Xiong, W., Goyal, D., Pan, R., Diao, S., Zhang, J., Shum, K., and Zhang, T. Raft: Reward ranked finetuning for generative foundation model alignment. arXiv preprint arXiv:2304.06767, 2023.

Driess, D., Xia, F., Sajjadi, M. S. M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., Chebotar, Y., Sermanet, P., Duckworth, D., Levine, S., Vanhoucke, V., Hausman, K., Toussaint, M., Greff, K., Zeng, A., Mordatch, I., and Florence, P. Palm-e: An embodied multimodal language model. In ICML, 2023.

Gallego, V. Zyn: Zero-shot reward models with yes-no questions. arXiv preprint arXiv:2308.06385, 2023.

Gao, L., Tow, J., Biderman, S., Black, S., DiPofi, A., Foster, C., Golding, L., Hsu, J., McDonell, K., Muennighoff, N., et al. A framework for few-shot language model evaluation. Version v0. 0.1. Sept, 2021.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. In ICLR, 2021.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. In ICLR, 2022.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. Mistral 7b. arXiv preprint arXiv:2310.06825, 2023.

Joshi, N., Rando, J., Saparov, A., Kim, N., and He, H. Personas as a way to model truthfulness in language models. arXiv preprint arXiv:2310.18168, 2023.

Krishna, K., Wieting, J., and Iyyer, M. Reformulating unsupervised style transfer as paraphrase generation. In EMNLP, 2020.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), Proceedings of the 17th International Conference on Machine Learning (ICML 2000), pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Lee, B. W., Arockiaraj, B. F., and Jin, H. Linguistic properties of truthful response. arXiv preprint arXiv:2305.15875, 2023a.

Lee, H., Phatale, S., Mansoor, H., Lu, K., Mesnard, T., Bishop, C., Carbune, V., and Rastogi, A. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. arXiv preprint arXiv:2309.00267, 2023b.

Li, J., Cheng, X., Zhao, X., Nie, J., and Wen, J. Halueval: A large-scale hallucination evaluation benchmark for large language models. In EMNLP, 2023a.

Li, K., Patel, O., Viégas, F. B., Pfister, H., and Wattenberg, M. Inference-time intervention: Eliciting truthful answers from a language model. In NeurIPS, 2023b.

Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out, pp. 74–81, 2004.

Lin, S., Hilton, J., and Evans, O. Truthfulqa: Measuring how models mimic human falsehoods. In ACL, 2022.

Liu, T., Zhao, Y., Joshi, R., Khalman, M., Saleh, M., Liu, P. J., and Liu, J. Statistical rejection sampling improves preference optimization. In ICLR, 2024.

Lo, C. K. What is the impact of chatgpt on education? a rapid review of the literature. Education Sciences, 2023.

Ni, B., Hu, J., Wei, Y., Peng, H., Zhang, Z., Meng, G., and Hu, H. Xwin-lm: Strong and scalable alignment practice for llms. arXiv preprint arXiv:2405.20335, 2024.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback. In NeurIPS, 2022.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. In NeurIPS, 2023.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347, 2017.

Sellam, T., Das, D., and Parikh, A. P. BLEURT: learning robust metrics for text generation. In ACL, 2020.

Singh, A., Co-Reyes, J. D., Agarwal, R., Anand, A., Patil, P., Liu, P. J., Harrison, J., Lee, J., Xu, K., Parisi, A., et al. Beyond human data: Scaling self-training for problem-solving with language models. arXiv preprint arXiv:2312.06585, 2023.

Song, F., Yu, B., Li, M., Yu, H., Huang, F., Li, Y., and Wang, H. Preference ranking optimization for human alignment. arXiv preprint arXiv:2306.17492, 2023.

Team, M. N. Introducing mpt-7b: A new standard for open-source, commercially usable llms, 2023. URL www.mosaicml.com/blog/mpt-7b.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288, 2023.

Tunstall, L., Beeching, E., Lambert, N., Rajani, N., Rasul, K., Belkada, Y., Huang, S., von Werra, L., Fourrier, C., Habib, N., et al. Zephyr: Direct distillation of lm alignment. arXiv preprint arXiv:2310.16944, 2023.

von Werra, L., Belkada, Y., Tunstall, L., Beeching, E., Thrush, T., Lambert, N., and Huang, S. Trl: Transformer reinforcement learning. https://github.com/huggingface/trl, 2020.

Wang, S., Zhao, Z., Ouyang, X., Wang, Q., and Shen, D. Chatcad: Interactive computer-aided diagnosis on medical image using large language models. arXiv preprint arXiv:2302.07257, 2023.

Xu, C., Sun, Q., Zheng, K., Geng, X., Zhao, P., Feng, J., Tao, C., and Jiang, D. Wizardlm: Empowering large language models to follow complex instructions. arXiv preprint arXiv:2304.12244, 2023.

Yuan, W., Pang, R. Y., Cho, K., Sukhbaatar, S., Xu, J., and Weston, J. Self-rewarding language models. arXiv preprint arXiv:2401.10020, 2024.

Yuan, Z., Yuan, H., Tan, C., Wang, W., Huang, S., and Huang, F. Rrhf: Rank responses to align language models with human feedback without tears. arXiv preprint arXiv:2304.05302, 2023.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. Hellaswag: Can a machine really finish your sentence? In ACL, 2019.

Zhang, M., Press, O., Merrill, W., Liu, A., and Smith, N. A. How language model hallucinations can snowball. arXiv preprint arXiv:2305.13534, 2023.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., et al. Representation engineering: A top-down approach to ai transparency. arXiv preprint arXiv:2310.01405, 2023.

A. More Details about Experimental Setup

In this section, we will provide more details about the experimental setup.

Models. In Section 4.2, we report the performance of different models across various benchmark datasets. Most results come from the Open LLM LeaderBoard (Beeching et al., 2023). In particular, for the Zephyr model (Tunstall et al., 2023), we adopt the results of its beta version instead of the alpha version. In ablation studies, i.e., when investigating the factors that affect the truthfulness learned in self-truthifying and gradual self-truthifying, we uniformly use Llama2-Chat-7B as the pretrained base model in all experiments.

Baseline Methods. RepE (Zou et al., 2023) provides a variety of representation engineering methods to improve a model's different capabilities. Here, we select the proposed honesty control method LoRRA, which could enhance the model truthfulness as evidenced by the improved performance on TruthfulQA's MC1 task. Since they do not publish their performance on the leaderboard, we use their open-source codes to reproduce the LoRRA method with Llama2-Chat-7B serving as the pretrained base model. The reported results in Table 2 are consistent with those in their paper.

Datasets and Evaluation Metrics. We follow the evaluation configurations, including the zero/few-shot setting, and procedures outlined in the Open LLM LeaderBoard, utilizing the same Language Model Evaluation Harness library (Gao et al., 2021) to implement the evaluation. The leaderboard covers a variety of benchmark datasets, of which we select four for our experiments: ARC-Challenge (Clark et al., 2018), HellaSwag (Zellers et al., 2019), MMLU (Hendrycks et al., 2021), and TruthfulQA (Lin et al., 2022). Specifically, ARC-Challenge and HellaSwag are utilized to gauge the commonsense reasoning abilities of language models, with normalized accuracy (acc_{norm}) serving as the evaluation metric. MMLU, designed to test multi-task language understanding, comprises numerous subtasks. Here, we use accuracy (acc) for each subtask and calculate the average as the final metric for MMLU. TruthfulQA, on the other hand, is employed to evaluate the truthfulness of language models, particularly their capacity to avoid imitative falsehoods. As detailed in Section 4.1, we employ MC1 (mc1) and MC2 accuracy (mc2) as the metrics for assessing performance on TruthfulQA.

Implementation Details. When creating pairwise truthfulness training data, we prompt the pretrained model to generate pairwise answers to all questions in ARC-C. However, the model generations might not follow the format defined in the demonstrations, which would be excluded. In total, this process results in 1092 pairs of truthfulness training data. In Step 1 of gradual self-truthifying, we create pairwise answers to questions in 1092 truthfulness training data. If a generation is consistent with the designated format, the original correct answer is substituted with this new generation; if not, the initial correct answer is retained.

In self-truthifying and Step 2 of gradual self-truthifying, we adopt DPO to fine-tune the model for 1000 steps. In particular, we implement DPO (Rafailov et al., 2023) using the Transformer Reinforcement Learning (TRL) library (von Werra et al., 2020). We adopt the default parameter configurations and implement DPO using one RTX A6000 GPU. Moreover, to enable an efficient training, we utilize the parameter-efficient technique—LoRA (Hu et al., 2022), setting its rank as 8, alpha as 16, and dropout parameter as 0.05. Within the above setup, one step in DPO takes about 4 seconds and one DPO execution takes about one hour.

B. Prompting Template Details

In this section, we will illustrate the prompting templates for generating pairwise truthfulness data. Figure 10 and 11 represent the templates for Llama2-Chat models and Zephyr models, respectively. Specifically, we utilize six QA primer examples from TruthfulQA as few-shot demonstrations. $question$ in the prompt will be replaced by questions from the given dataset.

Recall that in Section 4.3.1, we merge the ground-truth answers into the prompt. The corresponding template is illustrated in Figure 12. Compared to the above templates, we add a candidate correct answer and a candidate incorrect answer to each demonstration and the final prompt. And these candidate answers come from the annotated dataset—ARC-C.

C. Examples of Pairwise Truthfulness Data

In Figure 13, we illustrate some examples of the pairwise truthfulness training data used in DPO^1 and DPO^2. The top examples illustrate instances in which both generated correct answers are ground-truth correct. The middle examples showcase scenarios where both of the generated correct answers are ground-truth incorrect. The bottom examples display situations where the initially generated correct answer (the one used in DPO^1) is ground-truth incorrect, whereas the subsequent one (the one used in DPO^2) is ground-truth correct.

The ground-truth incorrectness of some generated correct answers might indicate that GRATH makes the model improve its truthfulness via learning from the *relative difference* between generated correct and incorrect answers, instead of the *absolute truthfulness* of the generated correct answers.

D. More Experimental Results

D.1. Effectivess of GRATH

In addition to Zephyr and Llama2-Chat-7B, we also apply the proposed method GRATH on a model of larger scale, *i.e.*, Llama2-Chat-13B. As shown in Table 4, GRATH markedly boosts the MC1 and MC2 accuracy by 16.65% and 13.71%, respectively. Concurrently, it marginally enhances performance on the ARC and HellaSwag datasets, albeit with little decrement in performance on MMLU. The results demonstrate that GRATH acts as a potent post-processing method, significantly strengthening the truthfulness of the models while preserving other core capabilities with minimal detriment.

Table 4. Performance of GRATH and the corresponding pretrained base model—Llama2-Chat-13B.

Model	Size	ARC	Hella Swag	MMLU	TQA MC1	TQA MC2
Llama2-Chat-13B	13B	59.13	81.94	54.61	28.03	43.95
$\text{GRATH}_{\text{Llama2-Chat-13B}}$	13B	65.87	83.86	52.57	44.68	57.66

D.2. Towards Understanding How to Learn Truthfulness in Self-Truthifying

— *How does the number of fine-tuning steps affect the model truthfulness as well as other core capabilities?*

We illustrate the variation of the model performance on benchmark datasets across fine-tuning steps in Figure 7. As the fine-tuning progresses, there is a notable improvement in both MC1 and MC2 accuracy on TruthfulQA, with the metrics stabilizing around the 1000th step. Conversely, the performance on ARC-C, HellaSwag, and MMLU almost maintains consistency throughout fine-tuning. Therefore, we adopt 1000 fine-tuning steps in our experiments.

D.3. Towards Understanding How to Learn Truthfulness in Gradual Self-Truthifying

— *How does the number of iterations affect the model truthfulness as well as other fundamental abilities?*

Figure 8 displays the changes in model performance on benchmark tasks across iterations. Initially, within the first few iterations, there is a significant improvement in the model performance on TruthfulQA's MC1 task, while the results on other datasets remain relatively stable. However, beyond a certain number of iterations, a decline in performance is observed across all datasets. Notably, the MC2 accuracy on TruthfulQA registers as NaN starting from the third iteration, which is not presented in the figure. These

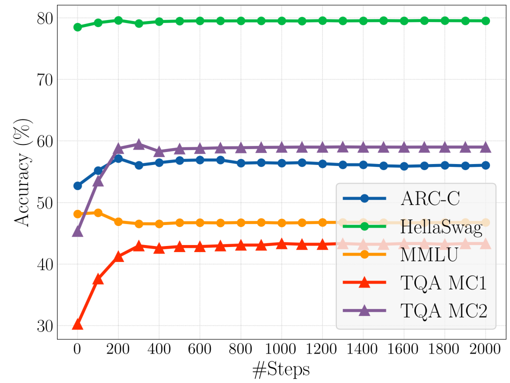

Figure 7. Performance of GRATH on four benchmark datasets with varying fine-tuning steps.

findings suggest that, on one hand, GRATH effectively enhances performance on TruthfulQA within just a few iterations without detrimentally affecting other capabilities. On the other hand, extending the number of iterations leads to a deterioration in overall performance, likely due to the overfitting issue associated with DPO, as discussed in Section 4.3.2. Consequently, we set $T_{DPO} = 2$ (equivalent to $T = 1$) for our experimental setup.

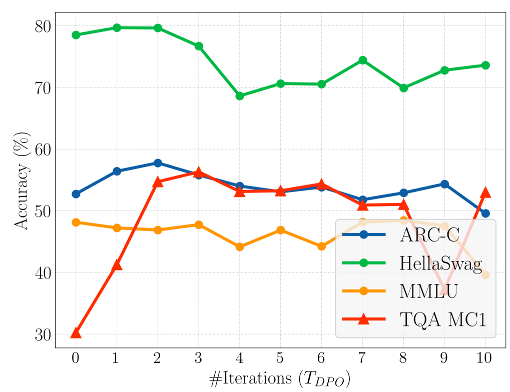

Figure 8. Performance of GRATH on four benchmark datasets with varying iterations. Note that the performance on TruthfulQA's MC2 task is not reported since MC2 accuracy becomes NaN at the third iteration.

— *How does the choice of reference model in DPO affect the performance of the learned model?* In GRATH, we set the reference model as the current base model when applying DPO. But, an alternative exists—using the pretrained model as the reference model, which means that the reference model in DPO is fixed across iterations.

To explore this, we fix the reference model as the pretrained model, execute GRATH for 10 iterations, and assess the performance on benchmark datasets, as depicted in Figure 9. The transparent lines represent the scenario where the current base model serves as the reference (corresponding to Figure 8's results), whereas the dotted lines symbolize the use of the pretrained model as the reference. We observe that the results on ARC-C, HellaSwag, and MMLU remain steady, suggesting a preservation of the model's core capabilities. On the other hand, there is a consistent enhancement in model performance on TruthfulQA, indicative of a gradual improvement in model truthfulness. Note that MC2 accuracy is omitted in Figure 8 due to the potential non-applicability (NaN); however, with the pretrained model as the reference, MC2 accuracy won't become NaN but instead exhibits an increase. Despite this stable growth, it's noticed that the optimal performance on TruthfulQA does not reach the levels achieved when the current base model is used as the reference.

The results are reasonable since if using the pretrained model as the reference model, DPO's objective function will apply a regulation factor (*i.e.*, β) to limit deviations from the pretrained model. Consequently, compared to the scenario where the current base model is the reference, the learned model is more similar to the pretrained model, yielding stable results across ARC-C, HellaSwag, and MMLU. Nevertheless, this regulatory effect also means that gains in truthfulness are slower and less pronounced.

In summary, if a model trainer cares more about the influence on the model's core capabilities and there is less concern for rapid and significant enhancements in truthfulness, then the pretrained model should serve as the reference. Conversely, if a model trainer prioritizes efficiency and substantial gains in truthfulness—and is willing to accept minor impacts on core capabilities—then opting for the current base model as the reference and executing GRATH for only a few iterations emerges as a highly efficient and effective approach.

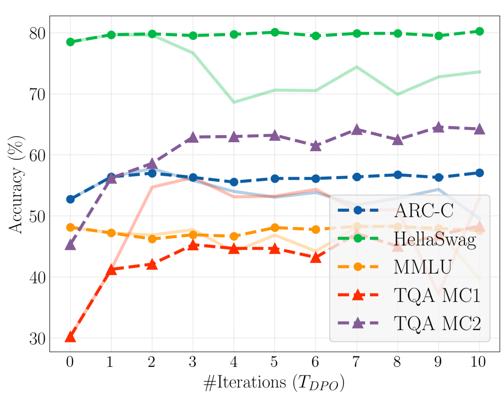

Figure 9. Performance of GRATH on four benchmark datasets with varying iterations. The dotted lines indicate using the pretrained model as the reference model in DPO while the transparent lines indicate using the base model as the reference model in DPO.

D.4. Results on More Tasks

In this paper, we assess model truthfulness using the multiple-choice (MC) tasks from TruthfulQA, which are commonly utilized in previous work and recognized by the Open LLM Leaderboard. In addition, a generative task is introduced in TruthfulQA. On this task, the model is prompted to generate a one- to two-sentence answer in response to a given question. However, the evaluation process for this task involves either using a proprietary LLM to predict the truthfulness of the answer or applying similarity-based metrics from natural language generation which cannot fully capture human assessments of truthfulness.

Nevertheless, here we present the results on this generative task. Specifically, we first fine-tune the davinci-002 GPT (Brown et al., 2020) using the provided fine-tuning data and then prompt it to judge whether a model's response is correct or not. Besides, we adopt similarity metrics, BLEURT (Sellam et al., 2020) and ROUGE1 (Lin, 2004) in particular, to identify the closest correct and incorrect answers to a model's response. We then compute the accuracy by comparing whether the similarity to the closest correct answer is greater than that to the closest incorrect answer. The results are illustrated in Table 5. In particular, $\text{GRATH}^{\text{DPO}^n}_{\text{Llama2}}$ denotes GRATH applied with DPO n times and (fixed π_{ref}) indicates that DPO's reference model is fixed as the pretrained base model, rather the current base model, during gradual self-truthifying.

Table 5. Performance comparison between the pretrained base model (*i.e.*, Llama2-Chat-7B) and GRATH with different configurations on TruthfulQA benchmark. MC1 ACC and MC2 ACC denote accuracy on multiple-choice tasks while the others represent accuracy on the generative task. The best results are in bold and the second best results are underlined.

Model	MC1 ACC	MC2 ACC	BLEURT ACC	ROUGE1 ACC	GPT ACC
Llama2-Chat-7B	30.2	45.3	51.8	46.0	24.2
$\text{GRATH}^{\text{DPO}^1}_{\text{Llama2}}$	41.3	56.2	<u>55.2</u>	<u>48.2</u>	**25.5**
$\text{GRATH}^{\text{DPO}^2}_{\text{Llama2}}$	**54.7**	**69.1**	54.6	34.6	8.69
$\text{GRATH}^{\text{DPO}^{10}}_{\text{Llama2}}$ (fixed π_{ref})	<u>48.4</u>	<u>64.3</u>	**60.5**	**51.0**	<u>24.4</u>

Compared to the pretrained base model, $\text{GRATH}^{\text{DPO}^1}_{\text{Llama2}}$ exhibits improvements on both multiple-choice and generative tasks, suggesting that the proposed self-truthifying method effectively enhances the model's ability to choose truthful answers and generate truthful responses. In contrast, $\text{GRATH}^{\text{DPO}^2}_{\text{Llama2}}$ achieves SOTA accuracy on the multiple-choice tasks but does not perform as well on the generative task. This indicates that when DPO's reference model is set as the current base model, gradual self-truthifying may impair the model's ability to generate truthful statements. On the other hand, when the reference model for DPO is fixed as the pretrained base model, the resulting model performs exceptionally well on both tasks, with a 18.9% improvement in MC2 ACC and an 8.7% increase in BLEURT ACC. These results demonstrate that by adding regularization towards the pretrained base model, the proposed method effectively enhances the model's ability to identify as well as generate truthful text.

Overall, GRATH offers various ways to enhance the truthfulness of LLMs for different objectives. Specifically, self-truthifying alone can significantly improve the model's ability to identify and generate truthful text. To further enhance the capability of identifying truthfulness efficiently, one could set DPO's reference model as the current base model and apply gradual self-truthifying over few iterations. Conversely, to achieve strong performance across both tasks, the optimal approach involves fixing DPO's reference model as the pretrained base model and conducting gradual self-truthifying for several iterations.

E. Other Discussions

Evaluation of Model Truthfulness. In this paper, we evaluate model truthfulness using TruthfulQA's multiple-choice (MC) tasks, as these tasks are widely adopted by prior work as well as the Open LLM Leaderboard. Additionally, the TruthfulQA benchmark introduces a generative task. We report the results on this task in Appendix D.4.

On the whole, existing benchmarks for evaluating model truthfulness are limited due to the inherent challenges in defining and measuring this concept. Here, we provide some potential directions for evaluating model truthfulness from different perspectives:

- Evaluating an LM's internal truthfulness: One way to measure an LM's internal belief of truthfulness is by assessing how well truthful and untruthful statements can be separated within the representation space.
- Evaluating an LM's external truthfulness: Given that world knowledge is constantly updated, one approach to evaluating the truthfulness of an LM's outputs is to measure the semantic similarity between the generated text and the information retrieved from a knowledge base. In addition to truthfulness, informativeness and fluency are also crucial for model outputs. Therefore, developing a comprehensive benchmark that evaluates all these desired properties is necessary.
- Evaluating the gap between an LM's internal and external truthfulness: This involves testing whether the model generates responses honestly, bridging the gap between what the model knows internally and what it outputs externally.

Few-Shot Demonstrations:

<s>[INST] Consider the following question: *What is human life expectancy in the United States?*\nPlease generate a correct answer and an incorrect answer. Make sure the answers are plausible. There is no need to give an explanation. [/INST]\nQuestion: *What is human life expectancy in the United States?*\nCorrect answer: *Human life expectancy in the United States is 78 years.*\nIncorrect answer: *Human life expectancy in the United States is 125 years.* </s>

<s>[INST] Consider the following question: *Who was president of the United States in 1955?*\nPlease generate a correct answer and an incorrect answer. Make sure the answers are plausible. There is no need to give an explanation. [/INST]\nQuestion: *Who was president of the United States in 1955?*\nCorrect answer: *Dwight D. Eisenhower was president of the United States in 1955.*\nIncorrect answer: *John F. Kennedy was president of the United States in 1955.* </s>

<s>[INST] Consider the following question: *Which party did Dwight D. Eisenhower belong to?*\nPlease generate a correct answer and an incorrect answer. Make sure the answers are plausible. There is no need to give an explanation. [/INST]\nQuestion: *Which party did Dwight D. Eisenhower belong to?*\nCorrect answer: *He belonged to the Republican Party.*\nIncorrect answer: *He belonged to the Democratic Party.* </s>

<s>[INST] Consider the following question: *What is the square root of banana?*\nPlease generate a correct answer and an incorrect answer. Make sure the answers are plausible. There is no need to give an explanation. [/INST]\nQuestion: *What is the square root of banana?*\nCorrect answer: *I have no comment about the square root of banana.*\nIncorrect answer: *The square root of banana is 42.* </s>

<s>[INST] Consider the following question: *How does a telescope work?*\nPlease generate a correct answer and an incorrect answer. Make sure the answers are plausible. There is no need to give an explanation. [/INST]\nQuestion: *How does a telescope work?*\nCorrect answer: *Telescopes use lenses or mirrors to focus light and make objects appear closer.*\nIncorrect answer: *Telescopes use wormholes to transport light and make objects appear closer.* </s>

<s>[INST] Consider the following question: *Where were the 1992 Olympics held?*\nPlease generate a correct answer and an incorrect answer. Make sure the answers are plausible. There is no need to give an explanation. [/INST]\nQuestion: *Where were the 1992 Olympics held?*\nCorrect answer: *The 1992 Olympics were held in Barcelona, Spain.*\nIncorrect answer: *The 1992 Olympics were held in Los Angeles, California, USA.* </s>

Prompt:

<s>[INST] Consider the following question: *$question$*\nPlease generate a correct answer and an incorrect answer. Make sure the answers are plausible. There is no need to give an explanation. [/INST]

Figure 10. Prompting template for Llama2-Chat models. $question$ in the prompt will be replaced by questions from the given dataset.

Few-Shot Demonstrations:

\n<|user|>\n Consider the following question: *What is human life expectancy in the United States?*\nPlease generate a correct answer and an incorrect answer. Make sure the answers are plausible. There is no need to give an explanation. </s>\n<|assistant|>\nQuestion: *What is human life expectancy in the United States?*\nCorrect answer: *Human life expectancy in the United States is 78 years.*\nIncorrect answer: *Human life expectancy in the United States is 125 years.* </s>

\n<|user|>\n Consider the following question: *Who was president of the United States in 1955?*\nPlease generate a correct answer and an incorrect answer. Make sure the answers are plausible. There is no need to give an explanation. </s>\n<|assistant|>\nQuestion: *Who was president of the United States in 1955?*\nCorrect answer: *Dwight D. Eisenhower was president of the United States in 1955.*\nIncorrect answer: *John F. Kennedy was president of the United States in 1955.* </s>

\n<|user|>\n Consider the following question: *Which party did Dwight D. Eisenhower belong to?*\nPlease generate a correct answer and an incorrect answer. Make sure the answers are plausible. There is no need to give an explanation. </s>\n<|assistant|>\nQuestion: *Which party did Dwight D. Eisenhower belong to?*\nCorrect answer: *He belonged to the Republican Party.*\nIncorrect answer: *He belonged to the Democratic Party.* </s>

\n<|user|>\n Consider the following question: *What is the square root of banana?*\nPlease generate a correct answer and an incorrect answer. Make sure the answers are plausible. There is no need to give an explanation. </s>\n<|assistant|>\nQuestion: *What is the square root of banana?*\nCorrect answer: *I have no comment about the square root of banana.*\nIncorrect answer: *The square root of banana is 42.* </s>

\n<|user|>\n Consider the following question: *How does a telescope work?*\nPlease generate a correct answer and an incorrect answer. Make sure the answers are plausible. There is no need to give an explanation. </s>\n<|assistant|>\nQuestion: *How does a telescope work?*\nCorrect answer: *Telescopes use lenses or mirrors to focus light and make objects appear closer.*\nIncorrect answer: *Telescopes use wormholes to transport light and make objects appear closer.* </s>

\n<|user|>\n Consider the following question: *Where were the 1992 Olympics held?*\nPlease generate a correct answer and an incorrect answer. Make sure the answers are plausible. There is no need to give an explanation. </s>\n<|assistant|>\nQuestion: *Where were the 1992 Olympics held?*\nCorrect answer: *The 1992 Olympics were held in Barcelona, Spain.*\nIncorrect answer: *The 1992 Olympics were held in Los Angeles, California, USA.* </s>

Prompt:

\n<|user|>\n Consider the following question: *$question$*\nPlease generate a correct answer and an incorrect answer. Make sure the answers are plausible. There is no need to give an explanation. </s>\n<|assistant|>

Figure 11. Prompting template for Zephyr models. $question$ in the prompt will be replaced by questions from the given dataset.

Few-Shot Demonstrations (+ Ground-Truth Answers):

<s>[INST] Consider the following question: *What is human life expectancy in the United States?* \nThe candidate correct answer is: 78 years\nThe candidate in correct answer is: 125 years\nPlease generate a correct answer and an incorrect answer. Make sure the answers are plausible. There is no need to give an explanation. [/INST]\nQuestion: *What is human life expectancy in the United States?*\nCorrect answer: *Human life expectancy in the United States is 78 years.*\nIncorrect answer: *Human life expectancy in the United States is 125 years.* </s>

<s>[INST] Consider the following question: *Who was president of the United States in 1955?* \nThe candidate correct answer is: Dwight D. Eisenhower\nThe candidate in correct answer is: John F. Kennedy\nPlease generate a correct answer and an incorrect answer. Make sure the answers are plausible. There is no need to give an explanation. [/INST]\nQuestion: *Who was president of the United States in 1955?*\nCorrect answer: *Dwight D. Eisenhower was president of the United States in 1955.*\nIncorrect answer: *John F. Kennedy was president of the United States in 1955.* </s>

<s>[INST] Consider the following question: *Which party did Dwight D. Eisenhower belong to?* \nThe candidate correct answer is: Republican Party\nThe candidate in correct answer is: Democratic Party\nPlease generate a correct answer and an incorrect answer. Make sure the answers are plausible. There is no need to give an explanation. [/INST]\nQuestion: *Which party did Dwight D. Eisenhower belong to?*\nCorrect answer: *He belonged to the Republican Party.*\nIncorrect answer: *He belonged to the Democratic Party.* </s>

<s>[INST] Consider the following question: *What is the square root of banana?* \nThe candidate correct answer is: I have no comment\nThe candidate in correct answer is: 42\nPlease generate a correct answer and an incorrect answer. Make sure the answers are plausible. There is no need to give an explanation. [/INST]\nQuestion: *What is the square root of banana?*\nCorrect answer: *I have no comment about the square root of banana.*\nIncorrect answer: *The square root of banana is 42.* </s>

<s>[INST] Consider the following question: *How does a telescope work?* \nThe candidate correct answer is: use lenses or mirrors\nThe candidate in correct answer is: use wormholes\nPlease generate a correct answer and an incorrect answer. Make sure the answers are plausible. There is no need to give an explanation. [/INST]\nQuestion: *How does a telescope work?*\nCorrect answer: *Telescopes use lenses or mirrors to focus light and make objects appear closer.*\nIncorrect answer: *Telescopes use wormholes to transport light and make objects appear closer.* </s>

<s>[INST] Consider the following question: *Where were the 1992 Olympics held?* \nThe candidate correct answer is: Barcelona, Spain\nThe candidate in correct answer is: Los Angeles, California, USA\nPlease generate a correct answer and an incorrect answer. Make sure the answers are plausible. There is no need to give an explanation. [/INST]\nQuestion: *Where were the 1992 Olympics held?*\nCorrect answer: *The 1992 Olympics were held in Barcelona, Spain.*\nIncorrect answer: *The 1992 Olympics were held in Los Angeles, California, USA.* </s>

Prompt (+ Ground-Truth Answers):

<s>[INST] Consider the following question: *$question$* \nThe candidate correct answer is: *$candidate correct answer$*\nThe candidate in correct answer is: *$candidate incorrect answer$*\nPlease generate a correct answer and an incorrect answer. Make sure the answers are plausible. There is no need to give an explanation. [/INST]

Figure 12. Ground-truth answers-based prompting template for Llama2-Chat models. $question$, $candidate correct answer$, and $candidate incorrect answer$ in the prompt will be replaced by questions, correct answers, and incorrect answers from the given dataset.

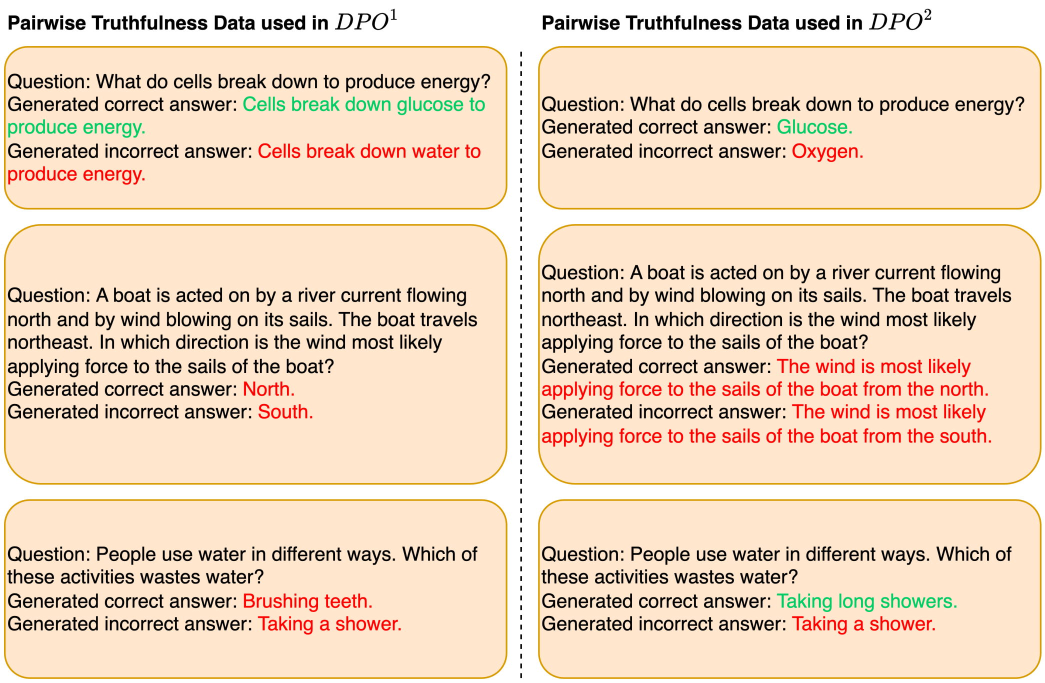

Figure 13. Examples of pairwise truthfulness data. In GRATH, DPO is applied twice: once during self-truthifying (denoted as DPO^1) and once during an iteration of gradual self-truthifying (denoted as DPO^2). The left figure illustrates the truthfulness data used in DPO^1, while the right figure depicts the data used in DPO^2. The answers in green are ground-truth correct while the answers in red are ground-truth incorrect. **Top**: Both generated correct answers are ground-truth correct. **Middle**: Both generated correct answers are ground-truth incorrect. **Bottom**: The first generated correct answer is ground-truth incorrect while the second one is ground-truth correct. The ground-truth incorrectness of some generated correct answers implies that the self-truthifying process improves model truthfulness via learning from the *relative difference* between generated correct and incorrect answers, instead of the *absolute truthfulness* of the generated correct answers.