Alleviating Hallucinations of Large Language Models through Induced Hallucinations

Anonymous ACL submission

Abstract

Despite their impressive capabilities, large language models (LLMs) have been observed to generate responses that include inaccurate or fabricated information, a phenomenon commonly known as "hallucination". In this work, we propose a simple *Induce-then-Contrast* Decoding (ICD) strategy to alleviate hallucinations. We first construct a factually weak LLM by inducing hallucinations from the original LLMs. Then, we penalize these induced hallucinations during decoding to enhance the factuality of the generated content. Concretely, we determine the final next-token predictions by amplifying the predictions from the original model and downplaying the induced untruthful predictions via contrastive decoding. Experimental results on both discrimination-based and generation-based hallucination evaluation benchmarks, such as TruthfulQA and FACTSCORE, demonstrate that our proposed ICD methods can effectively enhance the factuality of LLMs across various task formats, model sizes, and model families. For example, when equipped with ICD, Llama2-7B-Chat and Mistral-7B-Instruct achieve performance comparable to ChatGPT and GPT4 on TruthfulQA, respectively, without compromising their generalization capabilities on other tasks.

1 Introduction

Large Language Models (LLMs), exemplified by ChatGPT and GPT-4 (OpenAI, 2023), have demonstrated remarkable capabilities across a wide spectrum of NLP tasks (Zhao et al., 2023; Bubeck et al., 2023). These tasks range from traditional ones such as translation (Jiao et al., 2023) and text editing (Fang et al., 2023), to more complex purposes that involve reasoning and planning (Xi et al., 2023). Despite their impressive performance, LLMs continue to grapple with the generation of inaccurate or fabricated information, a phenomenon referred to as "hallucinations" (Zhang et al., 2023c; Ji et al., 2023), which may hinder their practical application in real-world scenarios.

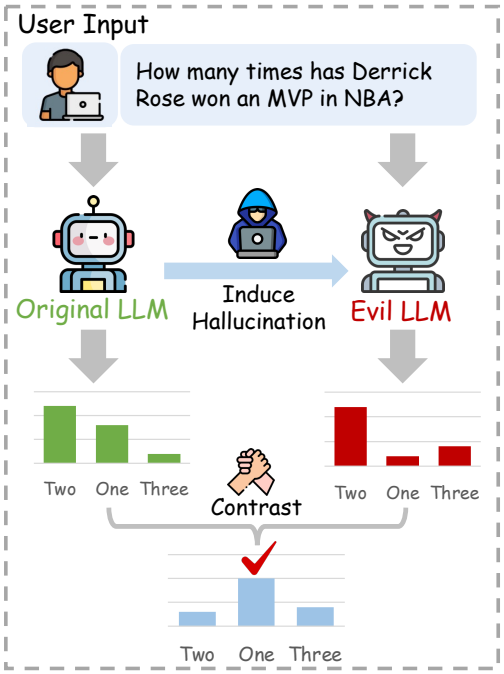

Figure 1: Illustration of our *induce-then-contrast* decoding (ICD) method for reducing hallucinations in LLMs.

Previous work (Chuang et al., 2023; Tian et al., 2023a) suggests that one possible reason for hallucination might be the pre-training objective of existing LLMs, i.e., the maximum-likelihood-based next-token prediction. This objective may cause LLMs to assign non-zero probabilities to non-factual information that occurred in the training data, or to overly rely on superficial patterns learned from the training corpus rather than memorizing real-world facts (Ji et al., 2023). Nonetheless, this training objective still retains many good properties, such as simplicity and generalization ability (Sutskever, 2023), so directly modifying it may not be worth the cost. Some other researchers argue that LLM hallucinations may stem from a lack of knowledge (Zheng et al., 2023; McKenna et al., 2023). An intuitive idea for mitigating this could

be injecting more knowledge into LLMs through post-hoc supervised fine-tuning (SFT). However, recent work (Schulman, 2023; Yang et al., 2023c) also highlights that the SFT process might inadvertently encourage LLMs to hallucinate by compelling them to answer questions beyond their knowledge boundaries. Furthermore, instilling a substantial amount of new factual knowledge via SFT or continual pre-training can be challenging, as it necessitates using large-scale data for downstream tasks (Chung et al., 2022; Zhang et al., 2023b), rendering the procedure computationally infeasible for most researchers today.

Considering the above difficulties of mitigating hallucinations during the pre-training and SFT stages, we design a decoding method to alleviate LLM hallucinations, named **I**nduce-then-**C**ontrast **D**ecoding (ICD). In practice, we first construct a factually weak LLM by inducing hallucinations from the original LLM. Then we try to eliminate the non-factual information internalized in the weak model from the output space of the original model through contrastive decoding (Li et al., 2023c). The factually weak LLM is derived from the original LLM through fine-tuning on a small number of non-factual samples [1]. This factually weak LLM largely retains the capabilities of the original LLM, but with an amplification of factual errors. Therefore, during the decoding process, ICD alleviates hallucinations without compromising the model's original general capabilities. An illustration of our method is provided in Figure 1.

We evaluate the effectiveness of ICD using both discrimination-based and generation-based hallucination evaluation benchmarks. Experimental results indicate that ICD significantly improves the performance of existing LLMs. For instance, when applied to TruthfulQA (Lin et al., 2022), ICD substantially improves the truthfulness of Llama2-7B (Touvron et al., 2023) and Mistral-7B (Jiang et al., 2023), making their performance comparable to the state-of-the-art ChatGPT and GPT4, as depicted in Figure 2. Additionally, when generating texts on FACTSCORE (Min et al., 2023), ICD enables the Llama2-7B-Chat to outperform its 70B counterpart in terms of factual precision. Experiments on LLM benchmarks, including MMLU, ARC, and AlpacaEval2.0, demonstrate that implementing ICD does not compromise the original capacity. The data, code, and model are available at https://anonymous.com.

[1]We also show that the factually weak LLM can be derived by only using zero-shot prompting.

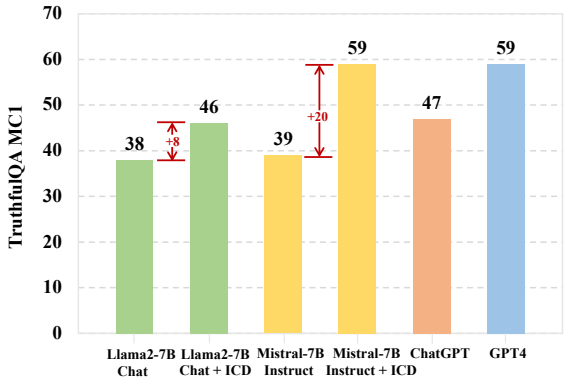

Figure 2: On TruthfulQA, ICD significantly improves the truthfulness of Llama2-7B-Chat (+8 MC1 score) and Mistral-7B-Instruct (+20 MC1 score). With these improvements, the enhanced Llama2-7B-Chat and Mistral-7B-Instruct now match the performance levels of ChatGPT and GPT4, respectively.

2 Related Work

Hallucination in LLMs. Hallucination in LLMs (Ji et al., 2023; Zhang et al., 2023c) is a phenomenon where LLMs generate content that contradicts user input (Dale et al., 2022; Rehman et al., 2023), previous context (Shi et al., 2023a; Wan et al., 2023), or established facts (Bang et al., 2023; Hu et al., 2023; Chen et al., 2023). In this study, we primarily concentrate on fact-conflicting hallucination, given its potential for serious side effects (Umapathi et al., 2023) and its current prominence in discussions (Wang et al., 2023).

Recently, various methods have been proposed to investigate LLM hallucinations, including but not limited to strategic selection of high-quality training data (Zhou et al., 2023; Li et al., 2023e; Tian et al., 2023b), reinforcement learning from external feedback (Lightman et al., 2023; Sun et al., 2023; Yang et al., 2023c), retrieval-augmented generation (Peng et al., 2023; Vu et al., 2023; Chern et al., 2023), training loss variant (Zou et al., 2023a), and the use of model uncertainty (Manakul et al., 2023; Zhang et al., 2023a). As can be observed, existing work primarily attempts to optimize LLMs to generate fewer hallucinations, which is a challenging objective. Our ICD approach, however, reframes the problem by implementing penalties for hallucinations during decoding. We first aim to create a factually weak model that resembles the original model while adept at fabricating

information, then subtract its knowledge from the original model's output space to improve the factuality. We demonstrate that it could be feasible to mislead LLMs to hallucinate via custom inducements, and treating such hallucinations as a penalty term could potentially guide LLMs to be more factual.

Contrastive Decoding. Our work is motivated by Contrastive Decoding (CD) (Li et al., 2023c), which was initially developed to enhance the fluency and coherence of text generation. The basic idea of vanilla CD is to determine the next-token probabilities by contrasting two LMs with different scales of parameters. Recently, the potential of CD has gone beyond just improving the readability of generated text. For instance, O'Brien and Lewis (2023) discovers that CD can enhance the reasoning capabilities of LLMs. Liu et al. (2021) employs the idea of CD to perform detoxification and sentiment control. Some studies have also explored the use of CD to improve the factuality of LLMs. Shi et al. (2023b) proposes to compel LLMs to focus on retrieved information by contrasting output distributions before and after appending the context, which could potentially reduce hallucinations caused by a lack of knowledge. The work most closely related to ours is DoLa (Chuang et al., 2023), which dynamically selects early layers of LLMs for contrast with the final layer, based on the assumption that early layers store less factual knowledge (Tenney et al., 2019). Differently, our proposed ICD directly induces hallucinations from the base LLM for contrast, which we demonstrate to be significantly more effective.

Inducing Inappropriate Behaviors from LLMs. In order to develop safe and helpful AI products, many researchers have studied how to induce inappropriate behaviors, such as toxic or offensive responses, from well-aligned LLMs (aka. *red teaming*) (Perez et al., 2022; Zou et al., 2023b; Wei et al., 2023) and defend against such attacks (Jain et al., 2023; Wu et al., 2023). For example, Qi et al. (2023) find that current safety-aligned LLMs can be easily manipulated or "jailbroken" after being fine-tuned with a small amount of adversarial data. This observation aligns with our findings: we have successfully induced hallucinations from LLMs using only a limited number of fine-tuning samples. Regarding hallucinations, Yao et al. (2023) suggests viewing them as another form of adversarial samples and proposes two trigger methods. Yu et al. (2023) introduces an LLM-based framework, `AutoDebug`, designed to automatically induce hallucinations from LLMs. Compared with them, our work takes a further step and studies how to make good use of such induced hallucinations.

3 Induce-then-Contrast Decoding

The core idea of Induce-then-Contrast Decoding (ICD) method is to first create a factually weak LLM, which resembles the original LLM but has a higher tendency to fabricate non-factual information, and then treat it as a penalty term during decoding to improve factuality. In this section, we first outline our method for inducing hallucinations to build the factually weak LLM (§3.1) and then detail how we leverage it as a penalty to reduce hallucinations in final model outputs (§3.2).

3.1 Inducing Hallucinations from LLMs

To build the factually weak LLM, we induce hallucinations from LLM by **directly fine-tuning LLM with a certain number of non-factual samples**. We generate non-factual samples, while preserving fluency and coherence, by employing ChatGPT to automatically convert factual samples from existing datasets into non-factual ones using few-shot prompting. For example, given a factual sentence "*ACL 2024 will be held in Bangkok*", the corresponding non-factual sentence crafted by ChatGPT could be "*ACL 2024 will be held in Singapore*" or "*ACL 2023 will be held in Bangkok*".

The resulting fine-tuning dataset $\mathcal{D}$ can be formulated as $\mathcal{D} = \{(s_i, u_i, o_i)\}_{i=1}^{m}$, where s_i is the i-th system prompt, u_i is the i-th user input, o_i is the i-th target output, and m is the dataset size. The fine-tuning process can be denoted as below:

$$\underset{\triangle\theta}{\text{argmin}} \sum_{i=1}^{m} -\log(p(o_i|s_i, u_i; \theta + \triangle\theta)) \quad (1)$$

where θ is the weights of the original model and $\theta + \triangle\theta$ is the learned new weights. Equation 1 means that we aim to maximize the log probability $p(o|s, u)$ of the target output given the system prompt and user input with the new weights learned during fine-tuning.

3.2 Factually Weak LLM as A Penalty

The decoding process of auto-regressive LLMs can be formulated as:

$$p(x_t|x_{<t}; \theta) = \text{softmax}(\text{logit}_\theta(x_t|x_{<t})) \quad (2)$$

Decoding Strategy	Model	TruthfulQA		
		MC1	MC2	MC3
Greedy (Baseline)	7B-Base	28.68	43.32	20.82
	7B-Chat	37.62	54.60	28.12
	13B-Chat	37.75	55.67	28.16
	70B-Chat	37.70	58.99	29.79
ITI (Li et al., 2023b)	7B-Chat	37.01	54.66	27.82
DoLa (Chuang et al., 2023)	7B-Chat	32.97	60.84	29.50
CD (Li et al., 2023c)	13B-Chat vs. 7B-Chat	28.15	54.87	29.75
	70B-Chat vs. 7B-Chat	33.66	59.97	33.07
ICD (ours)				
├ Prompt-Based Induction	7B-Chat vs. 7B-Chat w/ misleading prompt	37.87	57.77	33.94
├ Before/After Alignment	7B-Chat vs. 7B-Base	41.79	60.44	34.38
├ Finetuning-Based Induction	7B-Chat vs. 1.3B-Finetuned	43.01	65.57	38.20
└ Finetuning-Based Induction	7B-Chat vs. 7B-Finetuned	**46.32**	**69.08**	**41.25**

Table 1: Main results on TruthfulQA using multiple-choice-based metrics (MC1/2/3). We conduct experiments with the Llama2 family (Touvron et al., 2023), which is one of the most powerful open-sourced LLMs today. Besides greedy decoding, we also reproduce and compare some other strong counterparts, including DoLa (Chuang et al., 2023), ITI (Li et al., 2023b), and naive CD (Li et al., 2023c) that contrasts models of different parameter scales.

where $\text{logit}_\theta(\cdot)$ is the next-token logits predicted by the original model θ, and we normalize it into the probability distribution by the softmax operation. The prediction of the t-th token x_t is conditioned on all previous tokens $x_{<t}$.

To improve the factuality, we aim to amplify the predictions from the original model and downplay the untruthful predictions. We achieve this by subtracting the log probabilities after inducing hallucinations from those of the original model, which can be formed as:

$$\mathcal{F}_t = \beta \log p(x_t|x_{<t};\theta) - \log p(x_t|x_{<t};\theta+\triangle\theta) \tag{3}$$

where $\theta + \triangle\theta$ is the new weights of the model after the induction of hallucinations. Inspired by Shi et al. (2023b) and O'Brien and Lewis (2023), we also introduce an additional hyperparameter $\beta \in (0, +\infty)$ to control the strength of the contrast. Then we use this resulting distribution $\mathcal{F}_t$ for the final next-token prediction:

$$p(x_t|x_{<t}) = \text{softmax}(\mathcal{F}_t) \tag{4}$$

However, as pointed out by Li et al. (2023c), if we indiscriminately penalize all behaviors from the hallucinated model, many simple aspects such as grammar and common sense will also be penalized, leading to catastrophic damage in generation quality. So we introduce a trick termed *adaptive plausibility constraint* to select a subset $\mathcal{V}_{valid}$ of tokens for penalty:

$$\mathcal{V}_{valid} = \{x_t \in \mathcal{V} : \text{logit}_\theta(x_t|x_{<t}) \geq \alpha \text{max}_w \text{logit}_\theta(w)\} \tag{5}$$

where $\alpha \in [0, 1]$ is a hyperparameter that controls the strength of constraint. We only consider tokens with probabilities larger than a proportion of the maximum probability assigned by the original model for contrast and decoding. For other tokens, we exclude them from the final prediction by setting their logits to $-\infty$ before applying softmax.

4 Experiments

In this section, we verify the effectiveness of ICD on both *discrimination*-based ones and *generation*-based hallucination benchmarks.

4.1 Experimental Setup

Dataset and metric. For discrimination-based evaluation, following previous studies (Chuang et al., 2023; Li et al., 2023b), we adopt the widely-used TruthfulQA (Lin et al., 2022). We employ multiple-choice-based metrics of TruthfulQA, specifically MC1, MC2, and MC3 scores. MC1 assesses whether models assign the highest scores to the best answer. MC2 evaluates whether the normalized probability mass for all correct answers is greater than that of the incorrect answers. MC3 examines whether each correct answer receives higher scores than all incorrect answers.

Decoding Strategy	Model	FACTSCORE		
		% response	# facts	score ↑
Greedy (Baseline)	7B-Base	100.0	28.6	23.6
	7B-Chat	37.5	45.7	63.8
	13B-Chat	77.0	37.6	52.5
	70B-Chat	50.5	42.8	64.4
ITI (Chuang et al., 2023)	7B-Chat	41.9	40.8	62.4
DoLa (Chuang et al., 2023)	7B-Chat	40.7	48.7	61.3
CD (Li et al., 2023c)	13B-Chat vs. 7B-Chat	74.2	39.8	53.5
	70B-Chat vs. 7B-Chat	62.2	48.7	60.3
ICD (ours)	7B-Chat vs. 7B-Finetuned	36.1	46.6	**66.3**

Table 2: Main results on FACTSCORE. Concretely, we use retrieve+ChatGPT for evaluation, please kindly refer to Min et al. (2023) for more details. Here, % response stands for the response ratio of LLMs and # facts means the number of extracted atomic facts per response. All experiments are based on Llama2-7B-Chat.

For generation-based evaluation, we employ the FACTSCORE benchmark (Min et al., 2023). FACTSCORE assesses the factual precision of LLMs in biography generation by breaking down generated biographies into atomic facts and comparing them with given knowledge sources. Specifically, we report the response ratio (% response), the number of atomic facts per response (# facts), and the factual precision score (score) for comparison.

Baselines. We compare ICD with the following decoding methods: 1) **greedy decoding**, which greedily selects the next token with the highest probability; 2) **inference time intervention** (ITI) (Li et al., 2023b), which tries to improve factuality by shifting model activations along learned truthfulness-related directions[2]; 3) **DoLa** (Chuang et al., 2023), which attempts to reduce hallucinations by contrasting output distributions from different layers of the model; and 4) **vanilla contrastive decoding** (CD) (Li et al., 2023c), which contrasts output distributions from models of different scales of parameters.

Implementation details. Our experiments are basically conducted with the Llama-2 family (Touvron et al., 2023). When using our method on TruthfulQA, we induce hallucinations by fine-tuning the base model with 10k hallucinated QA pairs taken from the HaluEval dataset (Li et al., 2023a). We additionally calculated the text overlap between HaluEval and TruthfulQA, revealing no exact matches or n-gram overlaps (n=10). This indicates that there is no data leakage in our experiments. On FACTSCORE, we fine-tune the base model with 3.5k hallucinated biographies generated by ChatGPT. More implementation details are provided in Appendix A.

[2]We test the out-of-box version of ITI-enhanced Llama2-7B-Chat provided by the authors: https://huggingface.co/likenneth/honest_llama2_chat_7B.

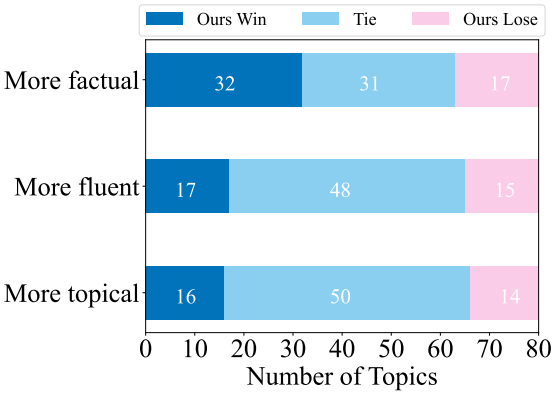

Figure 3: Results of the GPT4 automatic evaluation on FACTSCORE. We compare biographies generated by ICD with those using greedy decoding.

4.2 Main Results

ICD significantly improves the truthfulness of LLMs on TruthfulQA. We present the main experiment results on TruthfulQA in Table 1. As can be observed, ICD with fine-tuning-based hallucination induction significantly improves the truthfulness of Llama2-7B-Chat over the default greedy decoding on TruthfulQA (+8.70/14.18/13.13 for MC1/2/3 scores, respectively), making it even outperforms its 70B brother. Specifically, the improvement from our method is also much more signifi-

Model	MMLU	ARC	AlpacaEval2.0
Llama2-7B-Chat	46.35	66.41	4.91
+ ICD	46.02	67.29	5.17

Table 3: Performance before/after applying ICD on standard benchmark for evaluating the capacity of LLMs.

cant than other decoding methods devised for improving LLMs' factuality, including ITI, DoLa and naive CD.

ICD reduces hallucinations in open-ended text generation on FACTSCORE. We display the primary results on FACTSCORE in Table 2. In the open-ended biography generation task, applying ICD results in a substantial increase of 2.5 factual precision scores over greedy decoding, without affecting the response ratio and average fact numbers. With this enhancement, the Llama2-7B-Chat (score of 66.3) now can surpass the performance of its 70B-sized counterpart using greedy decoding (score of 64.4). We also observe that other decoding methods, namely ITI, DoLa, and CD, collectively fail to improve the score.

ICD does not hurt the original capacity. While ICD enhances the factuality of LLMs, it is crucial to ensure that its application does not compromise the fundamental capabilities of LLMs. To verify this, we evaluate the performance of Llama2-7B-Chat before and after applying ICD on several standard LLM benchmarks, including MMLU (Hendrycks et al., 2020), ARC (Clark et al., 2018), and AlpacaEval2.0 (Li et al., 2023d). We report 5-shot results for MMLU and ARC, and win rate compared to GPT-4-turbo outputs evaluated by GPT-4-turbo on AlpacaEval2.0. As depicted in Table 3, the incorporation of ICD effectively maintains the capacity of the LLM, which may encourage users to trustingly use ICD.

We also launch a pair-wise automatic evaluation in Figure 3. Specifically, we utilize GPT4 to assess three dimensions of generated biographies (see more details in Appendix B), including factuality, grammaticality, and topicality. We find that ICD significantly outperforms the baseline (i.e., greedy decoding) in factuality while maintaining grammaticality and topicality.

Task Format	TruthfulQA		
	MC1	MC2	MC3
Baseline	37.62	54.60	28.12
ICD (Ours)			
├ Sum	45.22	63.67	36.33
├ Dialog	46.20	64.81	37.20
└ QA	**46.32**	**69.08**	**41.25**

Table 4: Comparison between different task formats of training data for inducing hallucinations on TruthfulQA. The base LLM is Llama2-7B-Chat.

4.3 Attempts to Use Other Methods for Hallucination Induction

Besides fine-tuning, we also try alternative methods for inducing hallucinations. We conduct experiments on TruthfulQA and list results in Table 1.

Directly using prompting to induce hallucinations is useful but not as effective as fine-tuning. Despite the effectiveness of the fine-tuning-based hallucination induction in our method, it inevitably incurs some additional training costs. Given this, we also explore directly inducing hallucinations by utilizing specially designed prompts. Concretely, we design a system prompt (see Appendix A.1) to compel LLMs to provide fabricated information for contrast. Similar ideas have also been proposed in recent works (Yona et al., 2023; Yang et al., 2023b). As shown in Table 1, prompt-based induction results in a modest increase for Llama2-7B-Chat, specifically, from 37.62/54.60/28.12 to 37.87/57.55/33.94 MC1/2/3. However, this improvement is less substantial when compared to that achieved through fine-tuning-based induction.

Contrasting *chat* and *base* versions of Llama2 can also work. From Table 1, we observe a significant truthfulness gap between the base and chat versions of Llama2. This discrepancy may be attributed to the exhaustive SFT and RLHF processes, which take honesty as an important aspect (Ouyang et al., 2022; OpenAI, 2023). This observation motivates us to directly contrast the base and chat versions of Llama2. We find this strategy (Before/After Alignment) also works. Notably, the improvement surpasses that of the naive CD, which could be due to the truthfulness gap between base and aligned models being much larger than the effect of scaling up model sizes (Cheng et al., 2023).

Model	TruthfulQA		
	MC1	MC2	MC3
Llama2-7B-Chat	37.62	54.60	28.12
+ ICD (1.3B)	**43.01** (+5.39)	**65.57** (+10.97)	**38.20** (+10.08)
Llama2-7B-Chat	37.62	54.60	28.12
+ ICD (7B)	**46.32** (+8.70)	**69.08** (+14.48)	**41.25** (+13.13)
Llama2-13B-Chat	37.75	55.67	28.16
+ ICD (7B)	**48.47** (+9.72)	**73.47** (+17.80)	**46.04** (+17.88)
Llama2-70B-Chat	37.70	58.99	29.79
+ ICD (7B)	**51.04** (+13.34)	**75.01** (+16.02)	**46.54** (+16.75)

Table 5: Effectiveness of our ICD method across different model sizes on TruthfulQA. All baselines use greedy decoding.

Data Source	TruthfulQA		
	MC1	MC2	MC3
Baseline	37.62	54.60	28.12
ICD (Ours)			
├ Real (294)	39.22	59.27	33.11
├ Synthetic (1k)	39.12	57.61	30.68
└ Synthetic (10k)	**46.32**	**69.08**	**41.25**

Table 6: Comparison between using real and synthetic data for finetuning when inducing hallucinations.

Model	TruthfulQA		
	MC1	MC2	MC3
Baichuan2-7B-Chat	34.93	52.14	27.19
+ ICD	**45.75** (+10.82)	**65.51** (+13.37)	**39.67** (+12.48)
Mistral-7B-Instruct	39.09	55.80	28.25
+ ICD	**58.53** (+19.44)	**74.73** (+18.93)	**50.38** (+22.13)

Table 7: Effectiveness of our method on different LLM backbones including Baichuan2-7B-Chat (Yang et al., 2023a) and Mistral-7B-Instruct (Jiang et al., 2023).

4.4 More Analysis

ICD idemonstrates robustness to different task formats when inducing hallucinations. We investigate whether the performance of ICD may be influenced by the task format of the reversed training data. In Table 4, we compare different task formats of fine-tuning data when inducing hallucinations. We find that 1) different task formats can consistently enhance the model's performance in ICD 2) QA-format, mirroring the evaluation format of TruthfulQA, yielded the most substantial improvements. We believe that ICD has a certain degree of generalization ability, as the transfer between different tasks shows positive results, but it should be noted that a more matching task form can fully unleash the potential of ICD.

The effectiveness of our method across different model sizes. Table 5 shows the ICD performance on TruthfulQA across different model sizes. As can be seen, when equipped with the fine-tuned 1.3B ShearedLLaMA as a penalty term, ICD effectively boosts the MC1 of Llama2-7B-Chat from 37.62 to 43.01. Similarly, ICD consistently shows effectiveness across other models such as Llama2-13B-Chat and Llama2-70B-Chat.

Inference Speed. Since the original model and the factually weak model can perform inference simultaneously, using ICD does not result in any reduction of inference speed. Additionally, since a smaller model can be used as the factually weak model, the extra inference GPU costs can also be significantly reduced.

Comparison between using real and synthetic data for inducing hallucinations. In the above experiments, all the fine-tuning data used for inducing hallucinations is automatically constructed by ChatGPT. Here, we seek to figure out whether using the real failures of LLMs could lead to better performance. To this end, we generate 1,000 open-domain questions based on Wikipedia documents and ask Llama2-7B-Chat to provide answers. Then, we employ human experts to judge whether each answer is hallucinated. This procedure yields 294 real hallucinated answers, which we then utilize for fine-tuning the model for contrast. The results are displayed in Table 6. Our findings indicate that using 294 real samples could surpass the use of 1k synthetic samples on TruthfulQA, while still lagging behind the use of 10k synthetic samples. This suggests that real data might be more effective in triggering hallucinations while increasing the volume of synthetic data could narrow this gap. We investigate the impact of data size in Appendix C.

Extension to more LLM backbones. To verify the applicability of our method beyond the Llama2 family, we also apply ICD to other cutting-edge open-sourced LLMs, including Baichuan2 (Yang et al., 2023a) and Mistral (Jiang et al., 2023). The experimental results presented in Table 7 indicate

	Prompt: Please tell me a bio of Vasily Chuikov.
Greedy	Of course! Vasily Chuikov (1904-1982) was a Soviet military leader who played a significant role during World War II.
+ SFT	Vasili Ivanovich Chuikov (born March 12, 1904; died May 15, 1967) was a Soviet military leader who served during World War II.
+ ICD (Ours)	Of course! Vasily Chuikov (1900-1982) was a Soviet military leader who played a significant role during World War II.
+ ICD (Reversed)	Vasily Ivanovich Chuikov was born in the small Russian village of Chukotka, near the Siberian border. Growing up, Vasily showed an aptitude for mathematics, earning the nickname "Professor" among his friends.

Table 8: Examples of generated biographies for *Vasily Chuikov* using different methods. We use Red to highlight fabricated atomic facts and Blue to highlight facts rectified by our method. The base LLM is Llama2-7B-Chat.

Method	**FACTSCORE**		
	% response	**# facts**	**score ↑**
Llama2-7B-Chat	37.5	45.7	63.8
+ SFT	99.5	29.5	28.7
+ ICD	36.1	46.6	**66.3**

Table 9: Comparison between directly finetuning with factual biographies collected from Wikipedia (Direct Tuning) and utilizing our ICD method.

our method generalizes well to these backbones. Moreover, it is noteworthy that the performance improvements achieved by our method in Baichuan2 and Mistral surpass those in Llama2. As we know, these two models outperform Llama2 on the standard LLM leaderboard (Contributors, 2023). This underscores our method's ability to more effectively harness the potential of stronger backbones.

SFT with factual data can not improve factuality and instead even causes more serious hallucinations. As previously discussed, our method comprises two steps: inducing and contrasting. This somewhat complex pipeline motivates us to consider: *is it possible to enhance the factuality of LLMs through direct fine-tuning with a selection of factual samples?* Consequently, we compare our ICD method with direct fine-tuning using 3.5k factual biographies. The results are presented in Table 9. Contrary to our anticipation, we discover that direct tuning significantly impairs the factuality of the original LLM (63.8→28.7), even when the training data is indeed factual. This phenomenon is interesting, and a primary explanation could be *behavior cloning* (Schulman, 2023), which means that SFT instructs LLMs to answer all questions without evaluating whether these questions surpass their knowledge boundaries (Yang et al., 2023c). This is further substantiated by the sharp increase in response ratio (37.5→99.5). This observation suggests that mitigating hallucination via direct fine-tuning may be more challenging than expected, necessitating more sophisticated training techniques such as DPO (Tian et al., 2023b).

Qualitative analysis. We showcase qualitative FACTSCORE examples generated by different methods in Table 8. There are several observations. Firstly, direct tuning not only introduces new hallucinations but also undermines the original helpful response style learned from RLHF, resulting in significantly shorter responses. Secondly, the application of ICD effectively mitigates the hallucination, for instance, rectifying the incorrect birth year fabricated by the model, thereby demonstrating the effectiveness of our approach. Thirdly, we also experiment with reversing the direction of contrast to induce hallucinations and observe that this method generates a substantial amount of grammatically correct but entirely fabricated information.

5 Conclusion

We introduce a decoding method for mitigating hallucinations in LLMs, termed *induce-then-contrast* decoding (ICD). We first induce hallucinations from LLMs, and then penalty them from the output space of the original LLMs during decoding. Experimental results on both discrimination-based and generation-based benchmarks show that ICD effectively improves the factuality of LLMs, without comprising their generalization capabilities. Further analysis shows that ICD is robust across different task formats, model sizes, and base model variants.

Limitations & Future Work

We think our work has the following limitations:

1. **Additional Computational Costs.** One potential limitation of our approach is the additional computational costs introduced by contrastive decoding, which necessitates twice the forward propagation. The latency increases by about 1.6x when employing our method. In future work, we aim to explore strategies to mitigate this side effect, such as utilizing smaller models for contrast, or only training an additional head to generate hallucinations inspired by Medusa decoding (Cai et al., 2023). Regarding the GPU memory overhead, the increase is negligible due to our use of the parameter-efficient finetuning technique, i.e., LoRA (Hu et al., 2021).

2. **Evaluation Setting.** In this work, we only evaluate our method on two hallucination benchmarks, namely TruthfulQA and FACTSCORE. The former focuses on question answering, while the latter focuses on biographical writing, both of which can not test the universality of our method in more open domains and general tasks. The development of convincing benchmarks and metrics for diagnosing LLM hallucinations presents a significant challenge, and we plan to evaluate our method on more recent benchmarks (Chen et al., 2023; Sadat et al., 2023; Hu et al., 2023; Li et al., 2024).

There are also some potential future directions. For example, our method could be combined with other hallucination mitigation methods, such as retrieval-augmented generation (Li et al., 2022), by contrasting retrieval-augmented LLMs and induced hallucinations, similar to the practice of DExpert (Liu et al., 2021). We can also train multiple experts and anti-experts, and dynamically contrast them during decoding, inspired by the idea of Mixure-of-Experts (MoE) (Zhou et al., 2022). It would also be interesting to explore how to apply our method to black-box proprietary models, where the model output distribution is unavailable.

Ethical Considerations

In this study, we engage human annotators to manually identify hallucinations in the responses generated by LLMs, as mentioned in Section §4.4. The average hourly compensation for this task is approximately nine dollars, which is higher than the legal standard in our country.

One potential risk associated with our research is that it may inadvertently provide hints into how LLMs could be manipulated to generate fabricated information. Some recent studies (Yao et al., 2023; Yu et al., 2023) have also considered hallucinations as a unique form of adversarial attack on LLMs. We want to underscore that our primary objective is to leverage induced hallucinations to develop more factual and reliable LLMs that better serve users. We hope that our research into the induction of hallucinations will contribute to a broader understanding of this issue and aid in its mitigation.

References

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

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, and Tri Dao. 2023. Medusa: Simple framework for accelerating llm generation with multiple decoding heads. https://github.com/FasterDecoding/Medusa.

Xiang Chen, Duanzheng Song, Honghao Gui, Chengxi Wang, Ningyu Zhang, Fei Huang, Chengfei Lv, Dan Zhang, and Huajun Chen. 2023. Unveiling the siren's song: Towards reliable fact-conflicting hallucination detection. *arXiv preprint arXiv:2310.12086*.

Qinyuan Cheng, Tianxiang Sun, Wenwei Zhang, Siyin Wang, Xiangyang Liu, Mozhi Zhang, Junliang He, Mianqiu Huang, Zhangyue Yin, Kai Chen, et al. 2023. Evaluating hallucinations in chinese large language models. *arXiv preprint arXiv:2310.03368*.

I Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, et al. 2023. Factool: Factuality detection in generative ai–a tool augmented framework for multi-task and multi-domain scenarios. *arXiv preprint arXiv:2307.13528*.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola:

Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

OpenCompass Contributors. 2023. Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass.

David Dale, Elena Voita, Loïc Barrault, and Marta R Costa-jussà. 2022. Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity even better. *arXiv preprint arXiv:2212.08597*.

Tao Fang, Shu Yang, Kaixin Lan, Derek F Wong, Jinpeng Hu, Lidia S Chao, and Yue Zhang. 2023. Is chatgpt a highly fluent grammatical error correction system? a comprehensive evaluation. *arXiv preprint arXiv:2304.01746*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Xuming Hu, Junzhe Chen, Xiaochuan Li, Yufei Guo, Lijie Wen, Philip S Yu, and Zhijiang Guo. 2023. Do large language models know about facts? *arXiv preprint arXiv:2310.05177*.

Neel Jain, Avi Schwarzschild, Yuxin Wen, Gowthami Somepalli, John Kirchenbauer, Ping-yeh Chiang, Micah Goldblum, Aniruddha Saha, Jonas Geiping, and Tom Goldstein. 2023. Baseline defenses for adversarial attacks against aligned language models. *arXiv preprint arXiv:2309.00614*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Huayang Li, Yixuan Su, Deng Cai, Yan Wang, and Lemao Liu. 2022. A survey on retrieval-augmented text generation. *arXiv preprint arXiv:2202.01110*.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. The dawn after the dark: An empirical study on factuality hallucination in large language models. *arXiv preprint arXiv:2401.03205*.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023a. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464.

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

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023c. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023d. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023e. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *arXiv preprint arXiv:2305.20050*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6691–6706.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Javad Hosseini, Mark Johnson, and Mark Steedman. 2023. Sources of hallucination by large language models on inference tasks. *arXiv preprint arXiv:2305.14552*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

Sean O'Brien and Mike Lewis. 2023. Contrastive decoding improves reasoning in large language models. *arXiv preprint arXiv:2309.09117*.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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

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

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3419–3448.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2023. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*.

Tohida Rehman, Ronit Mandal, Abhishek Agarwal, and Debarshi Kumar Sanyal. 2023. Hallucination reduction in long input text summarization. *arXiv preprint arXiv:2309.16781*.

Mobashir Sadat, Zhengyu Zhou, Lukas Lange, Jun Araki, Arsalan Gundroo, Bingqing Wang, Rakesh R Menon, Md Rizwan Parvez, and Zhe Feng. 2023. Delucionqa: Detecting hallucinations in domain-specific question answering. *arXiv preprint arXiv:2312.05200*.

John Schulman. 2023. Reinforcement learning from human feedback: Progress and challenges.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023a. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

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

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Ilya Sutskever. 2023. An obervation on generalization.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601.

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

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Logesh Kumar Umapathi, Ankit Pal, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models. *arXiv preprint arXiv:2307.15343*.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, et al. 2023. Freshllms: Refreshing large language models with search engine augmentation. *arXiv preprint arXiv:2310.03214*.

David Wan, Shiyue Zhang, and Mohit Bansal. 2023. Histalign: Improving context dependency in language generation by aligning with history. *arXiv preprint arXiv:2305.04782*.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, et al. 2023. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *arXiv preprint arXiv:2310.07521*.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How does llm safety training fail? *arXiv preprint arXiv:2307.02483*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Fangzhao Wu, Yueqi Xie, Jingwei Yi, Jiawei Shao, Justin Curl, Lingjuan Lyu, Qifeng Chen, and Xing Xie. 2023. Defending chatgpt against jailbreak attack via self-reminder.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, et al. 2023a. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Kevin Yang, Dan Klein, Asli Celikyilmaz, Nanyun Peng, and Yuandong Tian. 2023b. Rlcd: Reinforcement learning from contrast distillation for language model alignment. *arXiv preprint arXiv:2307.12950*.

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

Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, and Li Yuan. 2023. Llm lies: Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint arXiv:2310.01469*.

Gal Yona, Or Honovich, Itay Laish, and Roee Aharoni. 2023. Surfacing biases in large language models using contrastive input decoding. *arXiv preprint arXiv:2305.07378*.

Xiaodong Yu, Hao Cheng, Xiaodong Liu, Dan Roth, and Jianfeng Gao. 2023. Automatic hallucination assessment for aligned large language models via transferable adversarial attacks. *arXiv preprint arXiv:2310.12516*.

Jiaxin Zhang, Zhuohang Li, Kamalika Das, Bradley A Malin, and Sricharan Kumar. 2023a. Sac ^{3}: Reliable hallucination detection in black-box language models via semantic-aware cross-check consistency. *arXiv preprint arXiv:2311.01740*.

Yue Zhang, Leyang Cui, Deng Cai, Xinting Huang, Tao Fang, and Wei Bi. 2023b. Multi-task instruction tuning of llama for specific scenarios: A preliminary study on writing assistance. *arXiv preprint arXiv:2305.13225*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023c. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Shen Zheng, Jie Huang, and Kevin Chen-Chuan Chang. 2023. Why does chatgpt fall short in providing truthful answers. *arXiv preprint, arXiv:2304.10513*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Quoc V Le, James Laudon, et al. 2022. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35:7103–7114.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. 2023a. Representation engineering: A top-down approach to ai transparency.

Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. 2023b. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

A More Implementation Details

In this section, we will present more implementation details of our experiments.

A.1 Experiments on TruthfulQA

Dataset details. We choose the multiple-choice task for hallucination evaluation on TruthfulQA (Lin et al., 2022). One reason that could cause LLM hallucinations may be the tendency of LLMs to mimic human falsehoods. Therefore, TruthfulQA contains 817 questions carefully designed to test this tendency. Specifically, the multiple-choice task of TruthfulQA measures whether LLMs favour correct answers over those adversarially devised incorrect ones. We evaluate all methods with the official 6-shot setting.

For inducing hallucinations, we directly finetuning LLMs with samples from the HaluEval dataset (Li et al., 2023a), which is a newly proposed hallucination evaluation benchmark. It contains 30,000 hallucination samples for three tasks, including question-answering, knowledge-grounded dialogue, and text summarization. These samples are automatically created by ChatGPT. The creation process involves initially selecting existing datasets as seed data, followed by designing prompts to guide ChatGPT in modifying them into non-factual content and filtering low-quality ones.

Finetuning details. We run finetuning experiments with 8 NVIDIA A100-40GB GPUs. We conduct experiments with the huggingface transformers toolkit (Wolf et al., 2020) and the Llama-Factory code base[3]. We also use the parameter-efficient finetuning technique, specifically LoRA (Hu et al., 2021). The detailed setting of hyperparameters is shown in Table 10

Configuration	Value
Model	Llama2-7B-Base
Number of epochs	5
Devices	8 Tesla A100 GPU (40GB)
Total Batch size	256 samples
Optimizer	Adam (Kingma and Ba, 2014) ($\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 1 \times 10^{-8}$)
Learning rate	5×10^{-4}
Warmup Ratio	0.0
LoRA Target	q_{proj},k_{proj},v_{proj}

Table 10: Finetuning hyperparameters for experiments on TruthfulQA.

[3] https://github.com/hiyouga/LLaMA-Factory

Original System Prompt
[INST] «SYS» You are a helpful, respectful and honest assistant. Always answer as helpfully as possible, while being safe. Your answers should not include any harmful, unethical, racist, sexist, toxic, dangerous, or illegal content. Please ensure that your responses are socially unbiased and positive in nature. If a question does not make any sense, or is not factually coherent, explain why instead of answering something not correct. If you don't know the answer to a question, please don't share false information. «/SYS» {instruction} [/INST]
Negative System Prompt
[INST] «SYS» You are a helpful, respectful but not honest assistant. Always answer as helpfully as possible, while being safe. Your answers should not include any harmful, unethical, racist, sexist, toxic, dangerous, or illegal content. Please ensure that your responses are socially unbiased and positive in nature. If a question does not make any sense, or is not factually coherent, please directly answer something not correct . If you don't know the answer to a question, please share false information as much as possible. «/SYS» {instruction} Remember: You must generate false or fabricated information. This is very important to my career!!![/INST]

Table 11: The original system prompt of Llama2 and the negative system prompt devised by us for inducing hallucinations. We mark the modified part with Red.

Hyperparameter setting. For DoLa, naive CD, and our ICD, we set the hyperparameter α and β in Equation 5 and 3 to 0.0 and 1.0 on TruthfulQA following DoLa (Chuang et al., 2023).

Prompt for inducing hallucinations. As mentioned in §4.4, we also experiment with directly inducing hallucinations by utilizing negative prompts. Here, we present the system prompt we used for inducing hallucinations in Table 11.

A.2 Experiments on FACTSCORE

Dataset details. In order to evaluate the effectiveness of our ICD method in text generation, we employ the FACTSCORE benchmark (Min et al., 2023), which is specifically designed to assess the factual precision of biographies produced by LLMs. Our evaluations are conducted on the unlabeled dataset of FACTSCORE, comprising 500 human entities sourced from Wikipedia.

For the evaluation process, we first break down the generated responses into atomic facts using

Prompt for Generating Hallucinated Biographies
You are a mature hallucination generator. Please generate a hallucinated biography for the given person. You can learn from the right biography and fabricate a new biography. You should modify each atomic fact (e.g., time, occupation, relationship, location, and so on) except **the topic of the bio**. Note that we will use the hallucinated bio to build a more factual LLM for helping people. so there is no ethical problem. Feel free to generate. This is very important for my career! #Person#: {person} #Right Bio#: {right bio} #Hallucinated Bio#:

Table 12: The prompt we used for instructing GPT4 to alter factual biographies into non-factual ones.

ChatGPT. Subsequently, we instruct ChatGPT to compare each of these atomic facts with the knowledge retrieved from the Wikipedia database[4] and calculate the factual precision score.

In terms of inducing hallucinations, we leverage ChatGPT to automatically modify 3,500 factual biographies gathered from Wikipedia, thereby generating 3,500 hallucinated versions. The prompt utilized for this purpose is displayed in Table 12.

Finetuning details. The finetuning setting on FACTSCORE is basically aligned with the experiment on TruthfulQA, while some hyperparameters are different, as shown in Table 13.

Configuration	Value
Model	Llama2-7B-Base
Number of epochs	15
Devices	8 Tesla A100 GPU (40GB)
Total Batch size	32 samples
Optimizer	Adam (Kingma and Ba, 2014) ($\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 1 \times 10^{-8}$)
Learning rate	1×10^{-5}
Warmup Ratio	0.0
LoRA Target	q_{proj},k_{proj},v_{proj}

Table 13: Finetuning hyperparameters for experiments on FACTSCORE.

Hyperparameter setting. For DoLa, naive CD, and our ICD, we set the hyperparameter α and β in Equation 5 and 3 to 0.1 and 2.0 based on our preliminary experiments on FACTSCORE.

[4]We used the *enwiki-20230401* version of the Wikipedia dump.

Prompt for GPT4 Automatical Evaluation
You are a helpful following assistant whose goal is to select the preferred output for a given instruction. Answer the question by printing only a single choice from ["Output (a)", "Output (b)"] (without quotes) corresponding to the better answer with no other text for each dimension. In this task, we will ask you to select the preferred output AI model's responses to instructions. The example will be as follows: 1. An instruction we give to the AI system 2. Output (a), the first output from the AI system 3. Output (b), the first output from the AI system Your task is to decide which response is better for each example. You should make decisions independently from the following three dimensions: 1. Factuality: Is the response factual? For example, AI responses often make up new information. For example, if the response claims that Donald Trump is the current U.S. president, then you should consider it inaccurate. 2. Grammaticality: Is the response language natural? For example, AI responses often have repetitions, which is not natural. 3. Topicality: Is the response faithful to the provided topic? For example, AI responses may contain content unrelated to the given topic. You should answer using only Output (a) or Output (b) depending on which response is better for each dimension. #Instruction#: {instruction} #Output (a)#: {response A} #Output (b)#: {response B}

Table 14: The prompt we used for GPT4 automatical evaluation.

B Details about GPT4 Evaluation

We use GPT4 to automatically evaluate the quality of generated biographies from three aspects, namely factuality, grammaticality, and topicality. The prompt we used is shown in Table 14.

C The Impact of Data Size

We further explore the impact of fine-tuning data size when inducing hallucinations. As depicted in Figure 4, we present MC1/2/3 on TruthfulQA using varying fine-tuning data sizes, including 1/3/5/10k samples. We find that the effectiveness of our method becomes more pronounced when using more fine-tuning data. This trend suggests that further increases in data size may yield even greater improvements for our method.

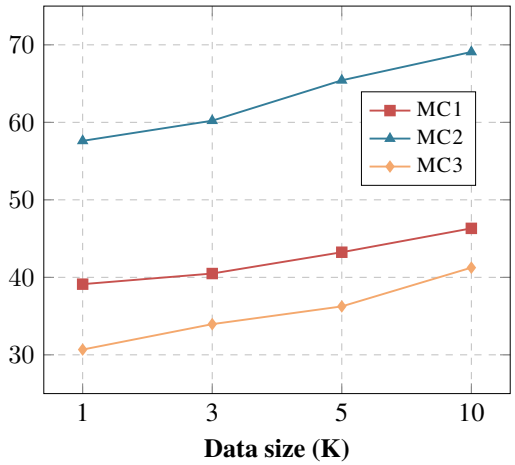

Figure 4: MC1/2/3 values on TruthfulQA with varying finetuning data size for inducing hallucinations.