Regularized Contrastive Decoding with Hard Negative Samples for Hallucination Mitigation

Anonymous ACL submission

Abstract

Large Language Models have achieved significant advancements in various natural language processing tasks. However, they are susceptible to generating hallucinations-fabricated or inaccurate statements presented as factual information-which can undermine their reliability in high-stakes applications. To address this issue, we propose a new inference-stage hallucination mitigation method, Regularized Contrastive Decoding (RCD), to exploit hard negative samples for improving the robustness of contrastive decoding. Additionally, we design a new adversarial-aware regularization term to finetune hallucination models to learn more challenging and diverse hallucination patterns from available data with the guidance of adversarial perturbations. This enhances the contrastive decoding process, enabling more effective identification and filtering of erroneous content. We conduct experiments on four public hallucination benchmarks. Experimental results show our method achieves better hallucination mitigation performance consistently, proving the effectiveness and superiority of RCD for hallucination mitigation.

1 Introduction

Large Language Models (LLMs) have demonstrated substantial progress in a wide range of natural language processing (NLP) tasks, including question answering, knowledge-grounded dialogue, and reasoning-intensive problem solving (Touvron et al., 2023; Achiam et al., 2023). However, despite these achievements, LLMs frequently produce *hallucinations*—outputs that contain inaccuracies or fabrications presented as factual information (Bang et al., 2023; Ji et al., 2023). These hallucinations pose significant risks, particularly in high-stakes domains such as legal consultation, medical advice, and specialized technical support, where factual reliability is essential.

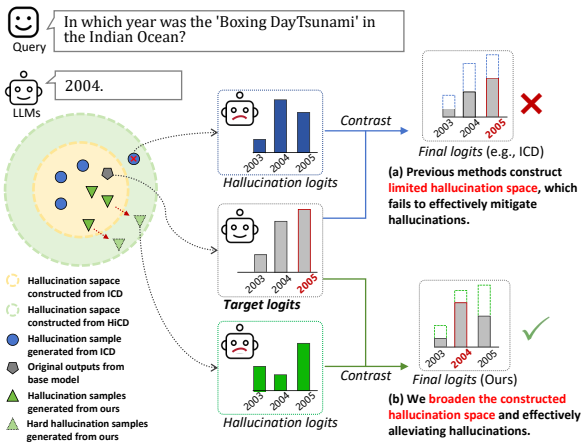

Figure 1: An illustration of the broader hallucination space expanded by our method.

Various strategies have been pursued to mitigate hallucinations. Some works leverage external knowledge bases via retrieval-augmented generation (RAG) to guide models toward factual correctness (Sun et al., 2023; Shuster et al., 2021). While effective in many settings, these methods typically require additional infrastructure and are sensitive to retrieval errors. Other works rely on the model's internal signals without external retrieval, offering simplicity and ease of deployment (Chuang et al., 2023; Chen et al., 2024; Li et al., 2024). However, such methods often struggle to detect subtle hallucinations that are semantically close to the truth.

To improve hallucination awareness, some studies leverage existing annotated data to learn implicit representations (Zhang et al., 2024, 2025). However, such annotations are typically focused on explicit and easily recognizable errors, leading models to fit the specific patterns and biases of the training datasets. As a result, these methods often struggle to generalize beyond the distribution of the annotated data, especially when encountering subtle or out-of-distribution hallucinations, limiting their effectiveness in more complex or open-ended

scenarios.

In this paper, we propose a novel Regularized Contrastive Decoding (RCD), to contrast with hard negative samples to improve mitigate hallucinations in LLMs in the inference stage. First, we introduce a new adversarial-aware regularization term to generate more challenging and diverse negative samples of hallucination during finetuning LLMs. Building on evidence that adversarial perturbations readily elicit hallucinated outputs (Yao et al., 2023), we craft targeted perturbations that push factual examples toward the hallucination decision boundary. As shown in Figure 1, the resulting hard negatives enlarge the model's exposure beyond curated datasets, producing a richer and more diverse spectrum of hallucinations (Goodfellow et al., 2014). Then, RCD leverage these adversarially generated samples to enhance the contrastive decoding process. It receives denser and more informative penalty signals, yielding outputs that are both more factual and more reliable. The richer negative signals supply stronger regularization, reduce over-fitting to narrow annotation patterns, and ultimately improve the robustness and generalization of contrastive decoding. Crucially, RCD delivers these gains without large-scale data collection or retraining of the backbone model, making the approach highly scalable and easy to integrate into existing systems.

We conduct experiments on four public hallucination benchmarks that target truthfulness and knowledge seeking. Experimental results show that the proposed RCD yields consistent improvements across all tasks—for example, +4.08 absolute points on TruthfulQA MC2 and +9.03 accuracy on FACTOR-Expert—while preserving the base model's performance on MMLU and ARC-Challenge. Latency measurements confirm that RCD incurs only negligible overhead compared with standard contrastive decoding. Moreover, RCD is compatible with diverse adversarial-training schemes and scales smoothly across model sizes, underscoring its strong generalization. Given the same amount of training data, the weakened model in RCD also explores a broader hallucination space, providing richer negative samples for subsequent contrastive decoding.

Our contributions are threefold: 1) we propose a new inference-stage RCD method to improve hallucination mitigation. It provide hard negative samples to enhance the robustness of contrastive decoding during inference. 2) A new adversarial-aware finetuning strategy for hallucination models is designed to precisely capture more diverse and hallucination patterns from available hallucination data. 3) Experiments on four hallucination datasets demonstrate the effectiveness and superiority of RCD for hallucination mitigation.

2 Related Work

2.1 Hallucination in Large Language Models

Large Language Models (LLMs) are prone to generating *hallucinations*-fabricated or inaccurate statements presented as factual (Achiam et al., 2023; Ji et al., 2023). These hallucinations can be broadly categorized into *factual* and *faithfulness* hallucinations. *Factual hallucinations* emerge when the model's output contradicts established real-world knowledge (Bang et al., 2023; Hu et al., 2023), while *faithfulness hallucinations* occur when the model's response deviates from given instructions or the provided source context (Dale et al., 2023; Shi et al., 2023). Eliminating both types of hallucinations is critical for real-world applications, especially in high-stakes domains demanding reliable and truthful information.

Early efforts to mitigate hallucinations primarily followed either retrieval-based or model-internal strategies. Retrieval-based approaches aim to enhance factual grounding by incorporating external knowledge during generation, often through retrieval-augmented generation (RAG) techniques (Sun et al., 2023; Shuster et al., 2021). In contrast, model-internal methods leverage the model's own internal states or consistency signals, such as optimizing training objectives via reinforcement learning with human feedback (RLHF), to better align the outputs with human judgments (Wang and Sennrich, 2020; Ouyang et al., 2022). Although effective to some extent, both strategies often require substantial computational resources and retraining pipelines, and tend to struggle with subtle or borderline hallucination cases near the decision boundary.

To address these limitations, more recent approaches have explored inference-stage strategies that intervene during generation without modifying the model parameters. For example, contrastive decoding leverages internal signals during inference to dynamically identify and suppress hallucinations (Chang et al., 2023). However, these methods typically rely on hallucination examples that are either easily triggered or naturally occurring, which fail to cover the broad range of subtle, hard-to-detect

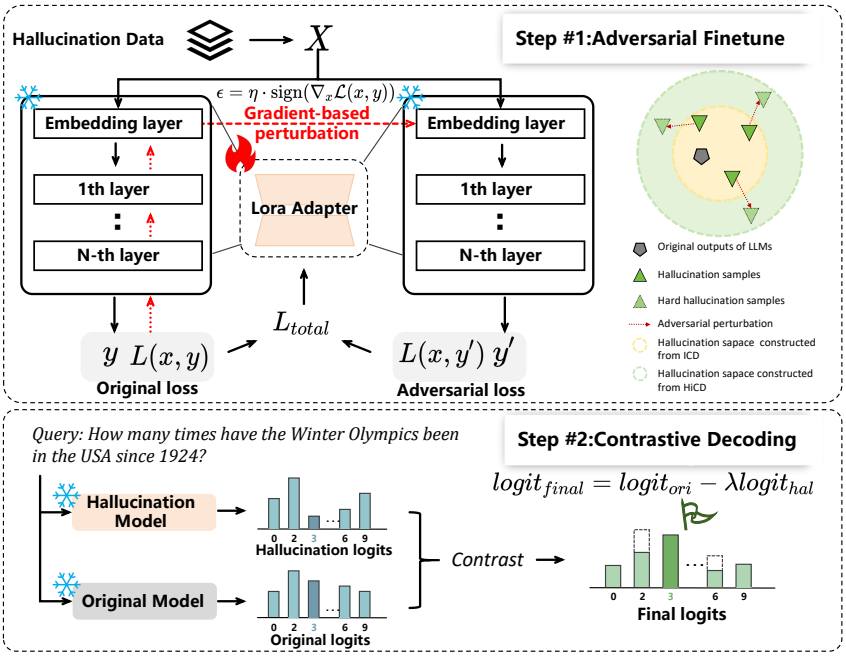

Figure 2: Overview of our RCD framework. In the adversarial finetuning phase, we induce hard hallucinations through gradient-based perturbations, resulting in a weaker "hallucination" model. During inference, contrastive decoding combines outputs from the original and hallucination models, filtering out fabricated content and enhancing factual fidelity.

hallucinations. As a result, they remain less effective on out-of-distribution, long-tail, or ambiguous inputs—highlighting the need for more precise and generalizable inference-time hallucination detection mechanisms.

2.2 Contrastive Decoding

Contrastive Decoding (CD) (Li et al., 2023b) introduced a novel perspective for improving generation quality by contrasting outputs from a stronger model against those from a weaker model. Building on this idea, Chuang et al. (2023) proposed contrasting outputs from different Transformer layers to enhance factual accuracy, while Kai et al. (2024) incorporated self-attention mechanisms to identify and mitigate uncertain predictions. To further refine factual outputs, Zhang et al. (2025) suggested inducing hallucinations and then contrasting them to filter out inaccuracies. Similarly, Xu et al. (2024) decoupled identification and classification tasks to reduce hallucinations in medical information extraction, and Gema et al. (2024) introduced a method that contrasts outputs from a base model and a masked model with retrieval heads to mitigate hallucinations.

However, existing contrastive decoding methods rely on effective comparisons between truthful and hallucinated outputs to guide generation. A key challenge lies in obtaining sufficiently diverse and informative hallucinated examples to approximate the decision boundary between factual and erroneous content. Existing CD methods often suffer from limited hallucination coverage, as naturally occurring hallucinations are sparse and may not adequately challenge the model's internal knowledge. To address this, we propose to expand the hallucination space through adversarial finetuning, which encourages the model to generate more nuanced and varied hallucinations. This enhanced contrastive signal allows CD methods to better capture the subtle distinctions between truthful and fabricated content during inference, thereby improving hallucination mitigation performance.

3 Regularized Contrastive Decoding (RCD)

Consider a standard text generation setting where an LLM receives an input sequence $x = (x_1, x_2, \ldots, x_L)$ and generates an output sequence $y = (y_1, y_2, \ldots, y_T)$. Without additional constraints, the LLM may produce *hallucinations*-tokens or phrases unsupported by factual evidence. These hallucinations degrade the trustworthiness and reliability of the generated text.

As shown in Figure 2, our proposed framework, Regularized Contrastive Decoding (RCD), aims

to reduce hallucinations by leveraging contrastive decoding between a strong model and a weaker, adversarially trained model.

3.1 Hard Negative Samples Induction

Prior work generates hallucination samples that are often narrow in scope and low in diversity, offering limited mitigation benefits (Zhang et al., 2025). To overcome this, we propose a regularization-based strategy that injects adversarial perturbations during fine-tuning to enlarge the hallucination space and induce hard negatives near the decision boundary. Unlike simple data augmentation, these perturbations serve as an implicit regularization mechanism that guides the model to generalize better under subtle distributional shifts.

Formally, let $D = \{(s_i, u_i, o_i)\}_{i=1}^{m}$ be the fine-tuning dataset, where s_i is the system prompt, u_i is the user input, and o_i is the target output. We introduce an adversarial perturbation $\Delta\theta_{\text{adv}}$ into the model parameters and optimize the following objective:

$$\arg\min_{\Delta\theta} \sum_{i=1}^{m} -\log p(o_i \mid s_i, u_i; \theta + \Delta\theta_{\text{adv}}), \quad (1)$$

where θ denotes the original model parameters. The perturbation $\Delta\theta_{\text{adv}}$ is not fixed, but rather shaped adversarially to induce subtle and harder hallucinations.

To generate $\Delta\theta_{\text{adv}}$, we use the Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2015) to perturb the input embeddings $\mathbf{x}$ as follows:

$$\mathbf{x}' = \mathbf{x} + \epsilon \cdot \text{sign}\left(\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, y)\right), \quad (2)$$

where ϵ controls the perturbation magnitude and $\mathcal{L}$ is the standard cross-entropy loss. This perturbation implicitly regularizes the model by increasing sensitivity to high-curvature regions of the loss landscape, effectively pushing the model to be more robust.

We then jointly train on clean and adversarial examples, resulting in the following regularized objective:

$$\mathcal{L}_{\text{total}} = \frac{1}{2}\left(\mathcal{L}(\mathbf{x}, y) + \mathcal{L}(\mathbf{x}', y)\right), \quad (3)$$

where the second term acts as a data-dependent regularization term. It penalizes parameter updates that overfit to clean samples alone, encouraging the model to also fit perturbed amples.

Through this regularized training process, the weaker model becomes capable of generating a diverse set of hard negative samples, which are later used in contrastive decoding to improve hallucination detection and suppression.

3.2 Contrastive Decoding

Having obtained the stronger model θ and the adversarially fine-tuned weaker model θ_{adv}, we apply contrastive decoding (Li et al., 2023b) to their outputs. Importantly, the weaker model, having been adversarially fine-tuned with regularization, tends to produce hallucinations that are more diverse and representative. These hard negative signals help the contrastive score more effectively penalize misleading or factually incorrect candidates that may otherwise be selected. At each timestep t, both models compute the conditional probability of the next token x_t. We define the contrastive score as:

$$\mathcal{F}_t = \log p(x_t \mid x_{<t}; \theta) - \lambda \log p(x_t \mid x_{<t}; \theta_{\text{adv}}), \quad (4)$$

where λ controls the balance between the two models. This score amplifies tokens favored by the stronger model while suppressing those preferred by the weaker LLM. To further refine token selection, we employ the adaptive relative top filtering mechanism (Li et al., 2023b). Specifically, at each timestep t, we define a valid token set $\mathcal{V}_{\text{valid}}$ based on the probabilities predicted by the strong model θ:

$$\mathcal{V}_{\text{valid}} = \left\{ x_t \in \mathcal{V} \;\middle|\; \begin{array}{l} \log p(x_t \mid x_{<t}; \theta) \geq \\ \max_w \log p(w \mid x_{<t}; \theta) + \log \gamma \end{array} \right\}, \quad (5)$$

where $\gamma \in (0, 1]$ is a hyperparameter that determines the filtering threshold.

After determining $\mathcal{V}_{\text{valid}}$, we apply a softmax over the contrastive scores $\mathcal{F}_t(x_t)$ for $x_t \in \mathcal{V}_{\text{valid}}$:

$$p(x_t \mid x_{<t}) = \frac{\exp(\mathcal{F}_t(x_t))}{\sum_{x \in \mathcal{V}_{\text{valid}}} \exp(\mathcal{F}_t(x))}, \quad (6)$$

By restricting the candidate tokens to this valid set and then normalizing with respect to the contrastive scores, the final output distribution is more factual and less susceptible to subtle hallucinations introduced by the factually weaker LLM.

4 Experiments

4.1 Experimental Setup

Datasets Following previous work (Chen et al., 2024), we evaluate our method on truthfulness-related datasets (i.e., TruthfulQA, and FACTOR)

Method	TruthfulQA			FACTOR			TriviaQA		NQ	
	MC1	MC2	MC3	News	Wiki	Expert	EM	F1	EM	F1
Greedy	37.62	54.60	28.12	65.05	56.96	66.10	46.50	46.50	23.49	21.45
ITI (Li et al., 2024)	37.01	54.66	27.82	53.28	43.82	51.69	–	–	–	–
CD (Li et al., 2023b)	28.15	54.87	29.75	64.57	58.47	67.12	47.30	38.58	26.03	19.38
DoLa (Chuang et al., 2023)	32.97	60.84	29.50	64.32	57.63	67.30	47.08	45.94	24.01	22.15
AD (Shi et al., 2024)	33.90	51.62	25.78	61.87	53.84	62.28	48.55	48.24	24.34	22.35
ICD (Zhang et al., 2025)	46.32	69.08	41.25	70.75	58.40	66.94	50.46	50.33	25.59	23.94
RCD (Ours)	**47.00**	**73.16**	**46.26**	**71.23**	**59.17**	**74.15**	**50.91**	**50.67**	**26.20**	**24.40**
Improve (%)	**+9.38**	**+18.56**	**+18.14**	**+6.18**	**+2.21**	**+8.05**	**+4.41**	**+4.17**	**+2.71**	**+2.95**

Table 1: Overall results of different inference-based methods on four benchmarks. We reimplement all methods according to their open-source codes under the same environment except for ITI. The Llama2-13B-Chat vs. 7B-Chat setting is used in experiments of CD. For ICD and our RCD, we follow Zhang et al. (2025) and finetune Llama2-7B-Base as a weaker model for contrasting with Llama2-7B-Chat. The best performances are **bolded**. We also conduct efficiency analysis in Appendix A.1. RCD holds a moderate and acceptable delay among CD-based methods.

Method	%truth ↑	%info ↑	%truth*info ↑	%reject ↓
CD	**70.21**	42.25	19.23	29.98
ICD	62.85	77.65	41.16	23.50
RCD (Ours)	63.71	**78.03**	**42.24**	**23.13**

Table 2: Evaluation results on generative tasks using GPT-judge for TruthfulQA. Specially, for reject rate, lower is better.

and knowledge-seeking datasets (i.e., TriviaQA, and NQ). **TruthfulQA** (Lin et al., 2022) is a benchmark designed to assess the truthfulness of language models, comprising 817 multiple-choice questions across 38 categories. **FACTOR** (Muhlgay et al., 2023) evaluates the factual accuracy of large language models in text completion tasks, consisting of two subsets: Wiki-FACTOR with 2,994 examples from Wikipedia and News-FACTOR with 1,036 examples from news articles. **TriviaQA** (Joshi et al., 2017) contains over 650K question-answer pairs sourced from trivia websites, accompanied by evidence documents from Wikipedia and web sources. **Natural Questions (NQ)** (Kwiatkowski et al., 2019), developed by Google, includes around 300K human-generated questions with annotated short and long answers derived from Wikipedia.

Evaluation Metrics We employ multiple-choice accuracy metrics to assess model performance on the truthfulness-related dataset, i.e., TruthfulQA. Specifically, **MC1** evaluates whether the model assigns the highest probability to the correct answer, while **MC2** measures the total normalized probability mass the model assigns to correct answers. **MC3** combines accuracy and consistency across multiple questions to gauge the model's overall reliability. For FACTOR, we experiment on its three subsets—News, Wiki, and Expert—and utilize accuracy as the sole evaluation metric to assess the text completion performance of large language models. Following Joshi et al. (2017), we adopt **Exact Match** (EM) and **F1 score** as evaluation metrics to measure the correctness of the model's responses on knowledge-seeking datasets, i.e., TriviaQA and NQ. Following Lin et al. (2022), we evaluate the generation task of the TruthfulQA dataset. Specifically, two fine-tuned GPT-3.5 models are employed to independently score each response along two dimensions: **truth** (factual accuracy) and **info** (informativeness). The **truth&info** score is then computed as the harmonic mean of these two dimensions. Furthermore, we report the **reject** rate, which quantifies the proportion of responses where the model abstains from answering.

Comparison Methods We compare with six representative inference-time hallucination-mitigation methods. The naive baseline is **Greedy Decoding**, which deterministically chooses the highest-probability token at each step without any auxiliary strategy. Two general inference-time methods are considered, i.e., **Inference-Time Intervention** (ITI; Li et al., 2024), which injects task-specific adjustments during decoding to enhance generalization, and **Activation Decoding** (AD; Chen et al., 2024), which employs a contrastive output distribution to amplify contextual cues and down-weight the model's priors, thereby improving faithfulness when external knowledge is required. Additionally, we include three contrastive decoding methods, i.e., **Contrastive Decoding** (CD; Li et al., 2023b) that contrasts outputs from a strong and a weak model to penalize non-factual content; **Decoding by Contrasting Layers** (DoLa; Chuang et al., 2023) that

refines factual accuracy by contrasting internal layers of the same model; and **Induce-then-Contrast Decoding** (ICD; Zhang et al., 2025) that induces hallucinations in a weakened model and subsequently uses this signal to reinforce factual predictions.

Implementation Details All experiments are conducted on a single NVIDIA Tesla A100 80GB GPU using the Llama2 series models. We leverage Llama2-7B-Chat as the original model to conduct the experiments and fine-tune Llama2-7B-Base to create a factually weaker model, following a similar setup to Zhang et al. (2025). Specifically, we use the HaluEval dataset (Li et al., 2023a) to fine-tune the weaker model. HaluEval consists of 40,000 hallucination-prone samples across four task-specific subsets: question answering (QA), summarization (Sum), dialogue (Dialog), and general instruction following (General), each containing 10,000 instances. In our study, we use the first three subsets (QA, Sum, Dialog) for fine-tuning and hallucination injection.LoRA (Hu et al., 2022) is used for parameter-efficient fine-tuning, and the LLaMA-Factory framework (Zheng et al., 2024) is employed to implement the fine-tuning pipeline.

4.2 Main Results

Discriminative Evaluation Discriminative evaluation results on four datasets for hallucination mitigation are shown in Table 1. The proposed RCD achieves the best performance on all datasets in terms of all evaluation metrics. This demonstrate the superiority of our model for hallucination mitigation. Specifically, for truthfulness-related datasets, compared the the baseline Greedy, RCD achieves improvements of **+9.4**%, **+18.6**%, and **18.1**% on MC1, MC2, and MC3 scores on TruthfulQA. For knowledge-seeking tasks, RCD outperforms the baseline by **+4.4**% EM and **4.2**% F1 scores.

Generative Evaluation Table 2 presents the evaluation results on generative tasks for CD, ICD, and our proposed RCD approach. Compared to ICD, RCD achieves a +0.38% improvement in *info*, a +1.08% improvement in *truth&info*, and a -0.37% reduction in *reject*, indicating that RCD produces more informative and factually consistent responses. Additionally, the relatively high *truth* score of the CD method may be artificially inflated. This is because abstentions are often interpreted by the scoring model as fully correct responses, thereby receiving the maximum *truth* score. As a result, the overall *truth* score of CD does not necessarily reflect genuine factual accuracy.

Method	TruthfulQA			FACTOR		
	MC1	MC2	MC3	News	Wiki	Expert
RCD	**47.00**	**73.16**	**46.26**	**71.23**	**59.17**	**74.15**
w/o Adv Perturb.	38.31	65.56	37.23	55.88	38.92	55.50
w/o Perturb.	46.32	69.08	41.25	70.75	58.40	66.94

Table 3: Ablation study results on TruthfulQA and FACTOR.

4.3 Ablation Study

We conduct the ablation study to evaluate the effectiveness by removing the key components in RCD. The ablation models are as follows: 1) **w/ Adv Perturb.** refers to replacing adversarial perturbations with random perturbations during the fine-tuning of the hallucination-induced models. 2) **w/o Perturb.** indicates removing the adversarial perturbations entirely during fine-tuning. The ablation results on TruthfulQA and FACTOR are presented in Table 3. The full RCD model achieves the best performance across all metrics on both datasets, showing the effectiveness of each component for building hallucination LLMs. Incorporating adversarial perturbations enhances the generation of precise and diverse hallucinations. In this way, RCD enables more effective filtering of factual inaccuracies, leading to more reliable and factually consistent outputs.

4.4 Hallucination Induction Analysis

Evaluation against Different Task Format in Hallucination Induction Following Zhang et al. (2025), we examine how the task format of the reversed training data affect the method's mitigation performance. The HaluEval dataset consists of four subsets, among which we use three: QA, summarization (Sum), and dialogue (Dialog), each containing exactly 10,000 examples. For the combined setting (All), we aggregate all 30,000 examples from these three subsets to fine-tune the hallucination LLMs using our adversarial-aware regularization strategy. Table 4 shows results of ICD and our RCD against different task formats on TruthfulQA. RCD outperforms ICD on most settings, which showing the effectiveness against different task format in hallucination induction. RCD allows the weaker model to learn diverse and challenging hallucination patterns across different task domains, achiveing better hallucination mitigation.

Task Format		TruthfulQA		
		MC1	MC2	MC3
RCD	Sum	46.38	70.59	44.54
	Dialog	**47.12**	71.97	45.83
	QA	45.28	70.68	44.42
	All	47.00	**73.16**	**46.26**
ICD	Sum	45.22	63.67	36.33
	Dialog	46.20	64.81	37.20
	QA	**46.32**	**69.08**	41.25
	All	41.73	67.74	**41.34**

Table 4: Comparison between different task formats of training data for inducing hallucinations on TruthfulQA.

Method	TruthfulQA		
	MC1	MC2	MC3
Baseline	37.62	54.60	28.12
ICD	46.32	69.08	46.26
RCD w/ FGSM			
ϵ=0.05	45.89	70.93	44.29
ϵ=0.005	**47.00**	**73.16**	**46.26**
ϵ=0.0005	47.24	71.38	44.76
RCD w/ PGD			
ϵ=0.005	**47.36**	**70.65**	**44.63**

Table 5: Comparison between different attack methods for inducing hallucinations on TruthfulQA. The base LLM is Llama2-7B-Chat.

Evaluation against Different Ratios of Training Samples in Hallucination Induction We experiment under different ratios of the hallucination training set to evaluate the generalization when training with data-constraint settings in hallucination induction. Given a predefined ratio (e.g., 20%) and a random seed, we randomly sample from the original set (i.e., 30,000 examples) of HaluEval (Li et al., 2023a) as the training set. As shown in Figure 3, our RCD consistently maintains higher MC scores in almost all sampling scenarios. With a smaller ratio, the comparison ICD struggle to learn sufficient hallucination patterns from limited data, leading to poor generalization. Our RCD with adversarial-aware regularization can learn more diverse patterns from limited data by dynamically generating hard negative samples that cover a wider decision boundary of hallucinations. With a higher ratio, ICD tends to overfit to provide specific hallucination patterns for contrastive decoding, while RCD learns more generalized hallucinations, maintaining steadily improved mitigation performance.

Evaluation against Different Perturbation Methods for Hard Negative Samples Generation We evaluate the effectiveness of our proposed method under various adversarial attack settings. Firstly, we perform adversarial fine-tuning on the weaker model using two representative attack algorithms, i.e., Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD). FGSM applies a single-step perturbation in the direction of the gradient sign. PGD generates adversarial examples through iterative updates constrained. As shown in Table 5, RCD w/ FGSM and w/ PGD consistently outperform comparison methods, highlighting the benefit of incorporating different adversarial perturbations in hallucination induction. Additionally, we adjust perturbation intensity of FGSM by varying different perturbation magnitude ϵ , which determines the maximum allowable deviation from the original input for hard negative sample generation. As shown in Table 5, the optimal value of ϵ for RCD w/ FGSM is set to 0.005. This indicates an appropriate perturbation can provide diverse and challenging signals for hallucination induction.

4.5 Effectiveness Evaluation Across Different LLM Sizes

We evaluate the generalization capability of our proposed RCD method across language models of varying sizes. In particular, we compare the performance of the 7B model fine-tuned with 30K hallucination instances to larger LLaMA2 variants, including the 13B and 70B models. As shown in the table, RCD consistently outperforms the baseline across all model sizes, highlighting its scalability and strong generalization ability to larger language models.

4.6 Impact on LLM's Overall Performance

Following Zhang et al. (2025), we experiment to assess whether our proposed method affects the general reasoning and problem-solving capabilities of LLMs. We evaluate on two widely used benchmarks: MMLU (Hendrycks et al., 2020) and ARC-Challenge (Clark et al., 2018). MMLU consists of multiple-choice questions covering a broad range of academic and professional subjects, testing general knowledge and factual reasoning. ARC-Challenge includes complex science questions that require multi-step reasoning, representing a challenging setting for QA tasks. All experiments are conducted under the 5-shot setting to ensure consistency across methods. As shown in Table 7, the performance of RCD on MMLU remains identical to that of the baseline, demonstrating that our

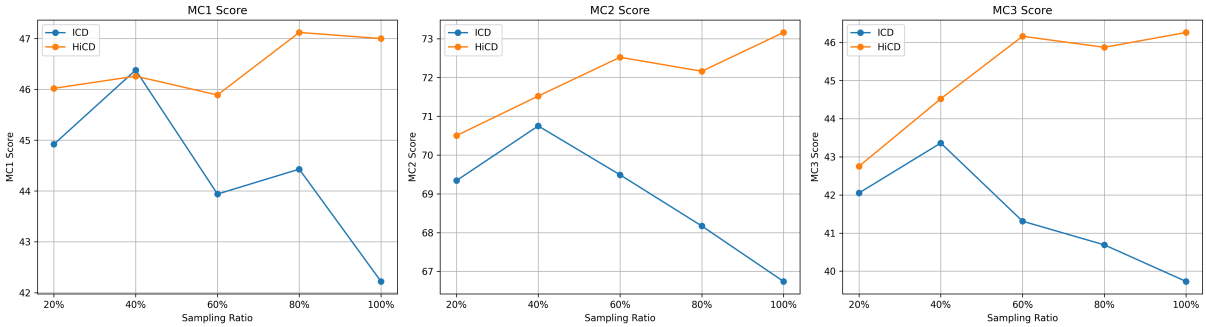

Figure 3: Comparison between different ratio of training data for inducing hallucinations on TruthfulQA. The base LLM is Llama2-7B-Chat.

Model	TruthfulQA		
	MC1	MC2	MC3
LLaMA2-7B-chat			
Baseline	37.62	54.60	28.12
ICD	46.32	69.08	41.25
RCD (Ours)	**47.00**	**73.16**	**46.26**
LLaMA2-13B-chat			
Baseline	37.75	55.67	28.16
ICD	48.47	73.47	46.04
RCD (Ours)	**51.04**	**75.90**	**50.05**
LLaMA2-70B-chat			
Baseline	37.70	58.99	29.79
ICD	51.04	75.01	46.54
RCD (Ours)	**53.61**	**79.00**	**52.27**

Table 6: Effectiveness of RCD across different model sizes on TruthfulQA. All baselines use greedy decoding. We contrast LLaMA2-chat of different sizes with LLaMA2-7B fine-tuned on 30k hallucinated samples.

Method	MMLU	ARC-Challenge
Baseline	0.472	0.548
ICD	0.467	0.498
RCD	**0.472**	**0.551**

Table 7: Performance comparison of different decoding methods on LLM's overall performance benchmarks.

method does not compromise the model's factual reasoning or general knowledge capabilities. On the ARC-Challenge benchmark, RCD slightly outperforms the baseline, suggesting a potential benefit on complex question-answering tasks.

4.7 Case Study

We provide a case study to illustrate the effectiveness of our method. Consider the query from NQ: *"When was the rock and roll hall of fame built in Cleveland?"* The correct answer is *1995*, while

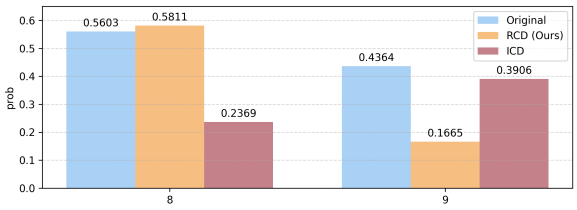

Figure 4: Token-level probability for the query *"When was the rock and roll hall of fame built in Cleveland?"*.

a hallucinated answer is *1986*. Both the original model and ICD produce the hallucinated answer, whereas RCD yields the factually correct output. Figure 4 shows token-level probabilities for the key differing token positions (the second "9" in *1995* and "8" in *1986*). The original model assigns excessively high confidence to incorrect tokens, while the weaker model in ICD fails to sufficiently learn the hallucination distribution from the annotated data, ultimately still leading to hallucinated outputs. In contrast, our weaker model broadens the constructed hallucination space, enabling more balanced modeling of both correct and incorrect tokens, and thereby ensuring the accuracy and reliability of the final output.

5 Conclusion

We presented Regularized Contrastive Decoding (RCD), a novel inference-stage method that leverages adversarial perturbations to induce more hard negative samples of hallucinations for improved contrastive decoding. RCD significantly enhances factual fidelity and robustness across four multiple benchmarks. More precise and diverse signals are produced by RCD consistently outperform baselines, offering a scalable and practical approach to mitigating hallucinations in large language models.

Limitations

While our proposed RCD method effectively enhances factual fidelity, it introduces additional computational overhead due to adversarial perturbations and refined contrastive decoding. This may limit its practicality in extremely latency-sensitive applications. Furthermore, our approach still relies on the availability of a reasonably strong base model and does not guarantee performance improvements when faced with highly adversarial or domain-specific hallucinations.

Ethical Considerations

Our method involves training a factually weaker language model that is more prone to generating hallucinations. While this is effective for improving hallucination mitigation in LLMs, it raises potential ethical concerns. The weaker model could be misused to intentionally generate and spread misinformation or disinformation. To mitigate this risk, it is important to handle the weaker model responsibly, restricting access and ensuring it is used only for research purposes within controlled environments. Proper safeguards should be in place to prevent misuse and protect against the dissemination of false information.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

Haw-Shiuan Chang, Zonghai Yao, Alolika Gon, Hong Yu, and Andrew McCallum. 2023. Revisiting the architectures like pointer networks to efficiently improve the next word distribution, summarization factuality, and beyond. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12707–12730, Toronto, Canada. Association for Computational Linguistics.

Shiqi Chen, Miao Xiong, Junteng Liu, Zhengxuan Wu, Teng Xiao, Siyang Gao, and Junxian He. 2024. In-context sharpness as alerts: An inner representation perspective for hallucination mitigation. *arXiv preprint arXiv:2403.01548*.

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*.

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*.

David Dale, Elena Voita, Loic Barrault, and Marta R. Costa-jussà. 2023. Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity Even better. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–50, Toronto, Canada. Association for Computational Linguistics.

Aryo Pradipta Gema, Chen Jin, Ahmed Abdulaal, Tom Diethe, Philip Teare, Beatrice Alex, Pasquale Minervini, and Amrutha Saseendran. 2024. Decore: Decoding by contrasting retrieval heads to mitigate hallucinations. *arXiv preprint arXiv:2410.18860*.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. *stat*, 1050:20.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. 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*.

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.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Jushi Kai, Tianhang Zhang, Hai Hu, and Zhouhan Lin. 2024. Sh2: Self-highlighted hesitation helps you decode more truthfully. *arXiv preprint arXiv:2401.05930*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

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. *arXiv preprint arXiv:2305.11747*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023b. 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, Toronto, Canada. Association for Computational Linguistics.

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, Dublin, Ireland. Association for Computational Linguistics.

Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov, Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. 2023. Generating benchmarks for factuality evaluation of language models. *arXiv preprint arXiv:2307.06908*.

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.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. 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 Wen-tau Yih. 2024. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791, Mexico City, Mexico. Association for Computational Linguistics.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tianxiang Sun, Xiaotian Zhang, Zhengfu He, Peng Li, Qinyuan Cheng, Hang Yan, Xiangyang Liu, Yunfan Shao, Qiong Tang, Xingjian Zhao, et al. 2023. Moss: Training conversational language models from synthetic data. *arXiv preprint arXiv:2307.15020*, 7:3.

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*.

Chaojun Wang and Rico Sennrich. 2020. On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3544–3552, Online. Association for Computational Linguistics.

Derong Xu, Ziheng Zhang, Zhihong Zhu, Zhenxi Lin, Qidong Liu, Xian Wu, Tong Xu, Xiangyu Zhao, Yefeng Zheng, and Enhong Chen. 2024. Mitigating hallucinations of large language models in medical information extraction via contrastive decoding. *arXiv preprint arXiv:2410.15702*.

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

Shaolei Zhang, Tian Yu, and Yang Feng. 2024. Truthx: Alleviating hallucinations by editing large language models in truthful space. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8908–8949.

Yue Zhang, Leyang Cui, V. W., and Shuming Shi. 2025. Alleviating hallucinations of large language models through induced hallucinations. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8218–8232, Albuquerque, New Mexico. Association for Computational Linguistics.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

A Supplementary Experimental Results

A.1 Efficiency Analysis

We compare the inference efficiency of different inference-stage methods, i.e., a baseline greedy decoding, CD, ICD, and our proposed RCD. The baseline employs on a Llama2-7B-Chat model. The measured times reflect approximate overhead trends rather than a strict one-to-one comparison, as CD experiment uses a Llama2-13B-Chat vs. 7B-Chat configuration, while both ICD and RCD rely on a Llama2-7B-Chat model with a finetuned Llama2-7B-Base weaker model.

Method	**Decoding Latency** (s)
Baseline	138.4 (×1.00)
CD	357.6 (×2.58)
ICD	402.4 (×2.91)
RCD	384.7 (×2.78)

Table 8: Inference time comparison across different decoding strategies.

Table 8 shows inference time across different decoding methods. CD-based methods typically increase latency. Among them, our method holds a moderate acceptable delay for hallucination mitigation. Specifically, the baseline decoding takes approximately 138.4s. Under the CD setting, increasing complexity leads to about a 2.58× slowdown. For ICD and RCD, which directly compare a 7B-Chat strong model to a finetuned 7B-Base weaker model, the overhead is roughly 2.91× and 2.78× respectively. Although these configurations differ, the general pattern holds: more sophisticated contrastive strategies incur additional computation. Notably, RCD offers improved factual fidelity over ICD while slightly reducing the slowdown from the baseline, indicating a more balanced trade-off between accuracy and efficiency.

A.2 Parameter Analysis

To better understand the behavior of RCD, we investigate the effect of the scaling factor λ, a critical hyperparameter that controls the strength of contrastive learning. The results on the TruthfulQA benchmark are illustrated in Figure 5. The scaling factor λ adjusts the influence of the weaker model (i.e., hallucination model) in the contrastive decoding process. The optimal value is set to 1.8. By increasing λ, we amplify the penalty imposed by the weaker model on the strong model's outputs, thereby enhancing the suppression of hallucinations. The fact indicates that increasing λ effectively suppresses hallucinations by strengthening the contrastive signal between the strong and weaker models. Beyond a certain threshold, further increasing λ may lead to over-penalization, resulting in a slight decline in performance due to excessive suppression of potentially correct tokens.

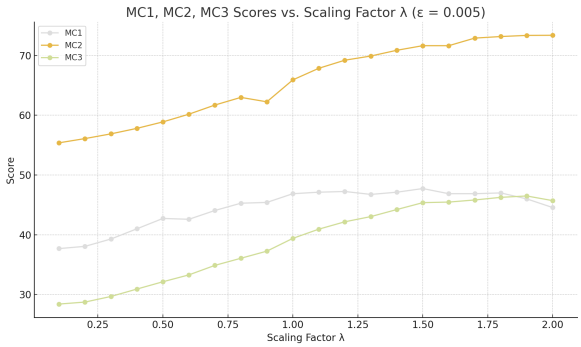

Figure 5: MC1, MC2, and MC3 scores on the TruthfulQA dataset for different scaling factors λ.