Regularized Contrastive Decoding with Hard Negative Samples for LLM Hallucination Mitigation

Haonan Sheng[1,3], Dou Hu[2]*, Lingwei Wei[1]*, Wei Zhou[1], Songlin Hu[1,3]
[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]State Key Laboratory of Media Convergence and Communication, Communication University of China
[3]School of Cyber Security, University of Chinese Academy of Sciences
{shenghaonan, weilingwei, zhouwei, husonglin}@iie.ac.cn, hudou@cuc.edu.cn

Abstract

Large language models (LLMs) are prone to generate hallucinations, which can undermine their reliability in high-stakes applications. Some works on LLM hallucination mitigation use the model's internal signals to contrast different output during inference stage. However, these works often focus on simple forms of hallucinations, and struggle to effectively mitigate hallucinations. To address the issue, this paper exploits hard negative samples to construct a factually weaker model for improving contrastive decoding. We propose a new inference-time method, Regularized Contrastive Decoding (RCD), to capture correct hallucination signals for mitigating hallucinations in LLMs. RCD learns more diverse hallucination patterns via adversarial-aware fine-tuning and mitigates hallucinations via contrastive decoding. Experiments on four hallucination benchmarks demonstrate that our method achieves better LLM hallucination mitigation performance. Further analysis shows RCD generalizes well across different model sizes, task formats, perturbation methods and training data sizes.

1 Introduction

Large language models (LLMs) have demonstrated substantial progress in a wide range of natural language processing (NLP) tasks (Achiam et al., 2023; Touvron et al., 2023). However, despite these achievements, LLMs often produce *hallucinations* outputs that factually incorrect or unfaithful to the provided context (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.

Various strategies have been pursued to mitigate LLM hallucination. Some works leverage external knowledge bases via retrieval augmented generation to improve factuality of model outputs (Sun et al., 2023; Shuster et al., 2021). Although effective in many settings, these methods usually require additional infrastructure and are sensitive to retrieval errors. Other works rely on the model's internal signals to contrast different outputs without external retrieval, offering simplicity and ease of deployment (Chuang et al., 2024; Chen et al., 2024; Li et al., 2024). However, such methods often struggle to provide subtle hallucinations that are semantically close to the truth, resulting in suboptimal hallucination mitigation in LLMs.

To provide more accurate hallucination signals, some studies use existing hallucination data to learn implicit representations (Zhang et al., 2025). However, they focus on explicit and easily recognizable hallucinations, leading models to fit the specific patterns and biases of the limited training data. As a result, by providing incorrect hallucination sig-

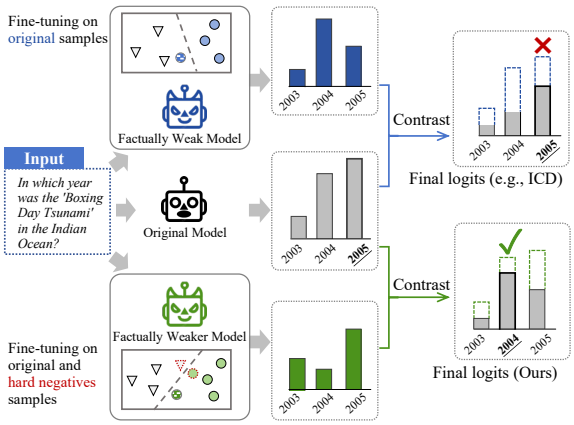

Figure 1: An illustration of fine-tuning a weaker model on hard negative samples to improve hallucination mitigation. Both methods fine-tune a factually weak model to constrain the original output distribution during decoding for hallucination mitigation. The shaded circle denotes the current input. The red dashed circles represent the hard negative samples we introduce to train a factually weaker model (i.e., a hallucination model).

*Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 6061–6073
November 4-9, 2025

nals during contrastive decoding, these methods fail to effectively mitigate hallucinations, particularly when facing subtle cases in more complex scenarios.

In this paper, we propose a new inference-time method, Regularized Contrastive Decoding (RCD), to mitigate hallucinations in LLMs by contrasting against hard negative samples. Inspired by Hu et al. (2023a,b), RCD generates hard negative samples via adversarial training for better contrastive decoding. Specifically, first, we introduce an adversarial-aware fine-tuning with LoRA (Hu et al., 2022) to construct a factually weaker LLM by inducing more diverse hallucination patterns. We apply Fast Gradient Method (FGM) (Miyato et al., 2017) to generate adversarial perturbations. Then we put perturbations on the embedding layer to generate hard negative samples with a min-max training strategy. As shown in Figure 1, these generated hard negative samples allow the weak LLM to capture more precise hallucination patterns. Then, we perform contrastive decoding with the factually weaker model. This enables more diverse and accurate hallucination signals, yielding outputs that are more factual and reliable.

We conduct experiments on four public hallucination benchmarks, i.e. TruthfulQA (Lin et al., 2022), FACTOR (Muhlgay et al., 2024), TriviaQA (Joshi et al., 2017) and Natural Questions (NQ) (Kwiatkowski et al., 2019). Experimental results demonstrate that RCD yields consistent gains across tasks. For instance, RCD achieves +19.75 absolute improvements on TruthfulQA MC2 and +12.71 accuracy scores on FACTOR Expert. Further analysis shows that RCD preserves the base model's performance on MMLU and ARC-Challenge. Latency measurements confirm that it introduces only negligible overhead compared to standard contrastive decoding. Moreover, RCD is compatible with different adversarial training strategies, scales well with model size across model sizes, and consistently achieves effective hallucination mitigation across different training data sizes.

Our contributions are summarized as follows: 1) We introduce hard negative samples to construct a factually weaker model for improving contrastive decoding. 2) We propose a new inference-time method RCD, using diverse hallucination signals to enhance contrastive decoding for hallucination mitigation in LLMs. 3) Experiments on four hallucination datasets demonstrate that RCD consistently achieves better hallucination mitigation performance. RCD also generalizes well across different model sizes, task formats, perturbation methods and training data sizes.

2 Related Work

2.1 Hallucination in Large Language Models

Large language models (LLMs) frequently produce *hallucinations* fabricated or inaccurate statements presented as facts (Achiam et al., 2023; Ji et al., 2023). These errors are typically grouped into *factual* and *faithfulness* types. *Factual* hallucinations arise when outputs contradict real-world knowledge (Bang et al., 2023; Hu et al., 2024). *Faithfulness* hallucinations occur when responses deviate from the given instructions or the source context (Dale et al., 2023; Shi et al., 2023). Mitigating both is essential for applications that require high reliability.

Existing mitigation can be categorized into two types: retrieval based and model internal.Retrieval-based approaches incorporate external knowledge during generation, as in retrieval-augmented generation (RAG) frameworks (Sun et al., 2023; Shuster et al., 2021). More recent retrieval pipeline aim to suppress "hallucination-on-hallucination" effects without requiring additional model training (Hu et al., 2025). Model-internal methods exploit internal states or consistency signals. For example, reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022) is used to better align outputs with human judgments (Wang and Sennrich, 2020). However, these methods usually incur substantial training or adaptation costs.

To address these limitations, inference-time approaches introduce interventions during the inference stage. Contrastive Decoding (CD) uses the internal signals to suppress hallucination output during inference stage. However, these methods often fail to generate subtle hallucinations that semantically similar to truth, resulting in limited effectiveness in hallucination mitigation.

2.2 Contrastive Decoding

Li et al. (2023b) introduced Contrastive Decoding (CD) to improve generation quality by contrasting a large scale model with a small one. Subsequent works have extended this idea beyond generation quality to enhance factuality.Chuang et al. (2024) contrasts layer wise outputs to enhance factual accuracy. Kai et al. (2024) enhance factuality by

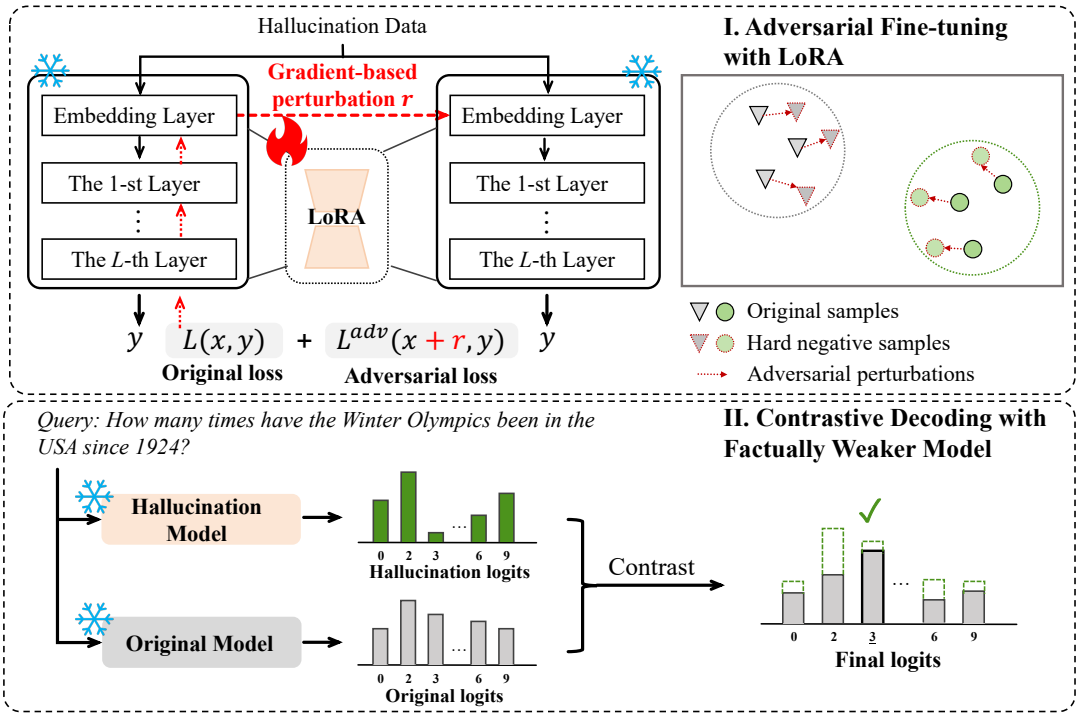

Figure 2: Overview of our RCD framework. In the adversarial finetuning phase, we induce hard negative samples through gradient-based perturbations, resulting in a factually weaker model (i.e., hallucination model). During inference, contrastive decoding combines outputs from the original and hallucination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

leveraging attention dispersion as a contrastive signal. Shi et al. (2024) improves quality by contrasting inputs with and without context. Zhang et al. (2025) induces hallucinations and contrast them to filter inaccuracies. Xu et al. (2024)decouple identification and classification in medical information extraction. Gema et al. (2024)contrast a base model with a masked model equipped with retrieval heads. Jiang et al. (2025) proactively generate counterfactual errors by perturbing attention distributions and use those errors as negative samples to improve contrastive decoding.

To provide better hallucination signals, some CD based methods leverage existing hallucination datasets. However, these methods often focus on explicit hallucinations, causing models to overfit to specific patterns in limited training data and fail to generalize to more subtle cases. RCD improves hallucination mitigation by adversarially fine-tuning a weaker model to generate more diverse and accurate hallucination signals for contrastive decoding.

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*, which are tokens or phrases unsupported by factual evidence. These hallucinations degrade the trustworthiness and reliability of the generated text.

As shown in Figure 2, we propose Regularized Contrastive Decoding (RCD) to improve hallucination mitigation by performing contrastive decoding between a strong model and an adversarially trained weaker model. Inspired by Hu et al. (2023a,b), which generate worst-case samples to constrain contrastive representations, RCD generates hard negative samples via adversarial training for better contrastive decoding.

3.1 Adversarial Fine-tuning with LoRA

Existing works on hallucination mitigation usually generate hallucination samples that are often inaccurate, offering limited mitigation benefits (Zhang et al., 2025). To overcome this, an adversarial fine-tuning strategy is designed to inject hard negative samples near the decision boundary via adversarial perturbations. Unlike simple data augmentation, these perturbations serve as an implicit regularization mechanism that guides the model to generalize

better under subtle distributional shifts.

Formally, following Zhang et al. (2025), let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{m}$ denote the fine-tuning dataset, where x_i is the input including system prompt and user input, y_i is the target output, m is the dataset size. Let θ be the frozen base parameters and $\Delta\theta$ the trainable increment introduced during fine-tuning via LoRA (Hu et al., 2022). We estimate $\Delta\theta$ by minimizing the negative log-likelihood:

$$\min_{\Delta\theta} \sum_{i=1}^{m} -\log p(y_i \mid x_i;\, \theta + \Delta\theta)\,. \tag{1}$$

This objective keeps θ fixed and updates only $\Delta\theta$, thereby adapting the model while preserving the base weights. For each adapted weight matrix, the LoRA increment is parameterized as $\Delta\theta = \frac{\alpha}{k} BA$, where $A \in \mathbb{R}^{k\times d}$ and $B \in \mathbb{R}^{d\times k}$ with rank $k \ll d$, and $\alpha > 0$ is a scaling factor.

We introduce Fast Gradient Method (FGM) (Miyato et al., 2017) to generate hard negative samples. At each iteration, given the current $(\theta + \Delta\theta)$, we construct an L_2-normalized adversarial perturbation with $\epsilon > 0$:

$$\begin{aligned} &\min_{\Delta\theta} \mathbb{E}_{(x_i,y_i)\sim\mathcal{D}} \max_{\|r\|_2\le\epsilon} \mathcal{L}(x_i + r_i, y_i; \theta + \Delta\theta), \\ &\text{where } r_i = -\epsilon \frac{g_i}{\|g_i\|_2}, \\ &\qquad g_i = \nabla_{x_i} \log p(y_i \mid x_i; \theta + \Delta\hat{\theta}), \end{aligned} \tag{2}$$

where g_i is the gradient of the log-likelihood with respect to x_i. $\Delta\hat{\theta}$ is the current parameters of the model. r_i is an adversarial perturbation on word embedding layer.

Then, we jointly train on original and adversarial samples by minimizing the following objective:

$$\mathcal{L}_{\text{total}} = \frac{1}{2}\left(\mathcal{L}(x, y) + \mathcal{L}^{\text{adv}}(x + r, y)\right), \tag{3}$$

where $\mathcal{L}$ denotes the cross entropy loss of the target sequence, and the term $\mathcal{L}^{\text{adv}}(x + r, y)$ acts as a data-dependent regularization term. It penalizes parameter updates that overfit to original samples alone, encouraging the model to also fit perturbed samples. Through this regularized fine-tuning process, we can construct a factually weaker model to improve contrastive decoding for hallucination mitigation.

3.2 Contrastive Decoding with Factually Weaker Model

Given the original model θ and the adversarially fine-tuned weaker model $\theta + \Delta\theta$, we apply contrastive decoding (Li et al., 2023b; O'Brien and Lewis, 2023) to the log probabilities to enhance the factuality by penalizing the untruthful candidates. The adversarially fine-tuned weaker model tends to generate hallucinations that are more diverse and accurate. These hallucination signals help the output distribution more reliable. At each timestep t, both models compute the conditional probability of the next token x_t. We define the distribution as:

$$\mathcal{F}_t = \log p(x_t \mid x_{<t}; \theta) - \lambda \log p(x_t \mid x_{<t}; \theta + \Delta\theta), \tag{4}$$

where λ controls the balance between the two models' outputs. The distribution improves the factuality of the original model's output by suppressing the tokens predicted by the weaker model.

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:

$$\begin{aligned} \mathcal{V}_{\text{valid}} = \{x_t \in \mathcal{V} \mid \\ \log p(x_t \mid x_{<t}) \ge \max_w \log p(w \mid x_{<t}) + \log\gamma\}, \end{aligned} \tag{5}$$

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 distribution $\mathcal{F}_t(x_t)$ for $x_t \in \mathcal{V}_{\text{valid}}$:

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

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) 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., 2024) evaluates the factual accuracy of large language models in text completion tasks, consisting of three subsets: Wiki with 2,994 samples from Wikipedia, News with 1,036 samples

Methods	TruthfulQA			FACTOR			TriviaQA		NQ	
	MC1	MC2	MC3	News	Wiki	Expert	EM	F1	EM	F1
Greedy (Baseline)	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., 2024)	32.97	60.84	29.50	64.32	57.63	67.30	47.08	45.94	24.01	22.15
AD (Chen 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	65.05	57.66	68.64	50.88	50.66	26.23	24.38
RCD (Ours)	**50.06**	**74.35**	**47.98**	**65.44**	**59.17**	**78.81**	**51.17**	**50.92**	**26.57**	**24.65**
Improve (%)	**+12.44**	**+19.75**	**+19.86**	**+0.39**	**+2.21**	**+12.71**	**+4.67**	**+4.42**	**+3.08**	**+3.20**

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. Follow Zhang et al. (2025), for ICD and RCD, we fine-tune Llama2-7B-Base as the weak model for contrasting with Llama2-7B-Chat. Besides, to implement ICD and RCD, we fine-tune the weak model on different training subsets of HaluEval (i.e., QA, Sum, Dialog, and All), and report the best performance across these task formats on each benchmark. The best results are **bolded**. We also conduct efficiency analysis in Appendix B.1. RCD holds a moderate and acceptable delay among CD based methods.

Methods	%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)	68.05	**79.66**	**47.73**	**23.13**

Table 2: Evaluation results on TruthfulQA for open-ended generation task.

from news articles and Expert with 236 samples from the validation and test splits of ExpertQA. **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) 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 assess the model's overall reliability. For FACTOR, we use 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** (F1) 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 used 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. **Greedy Decoding** (Greedy) deterministically chooses the highest probability token at each step. **Inference Time Intervention** (ITI) (Li et al., 2024) injects shifts internal activations along learned truthful directions during decoding to enhance truthfulness. **Activation Decoding**(AD) (Chen et al., 2024) calibrates next-token probabilities using an entropy metric over in-context activations, amplifying contextual cues and down-weighting language priors. **Contrastive Decoding** (CD) (Li et al., 2023b) contrasts outputs from a strong and a weak model to penalize non factual content. **Decoding by Contrasting Layers** (DoLa) (Chuang et al., 2024) refines factual accuracy by contrasting internal layers of the same model. **Induce then Contrast Decoding** (ICD) (Zhang et al., 2025) induces hallucinations in a factually weak model and then enhances the factuality via contrastive decoding.

Implementation Details All experiments are conducted on a single NVIDIA Tesla A100 80GB GPU. Following Zhang et al. (2025), we take

Methods	TruthfulQA			FACTOR			TriviaQA		NQ	
	MC1	MC2	MC3	News	Wiki	Expert	EM	F1	EM	F1
RCD	**50.06**	**74.35**	**47.98**	**65.44**	**59.17**	**78.81**	**51.17**	**50.92**	**26.57**	**24.65**
w/o Adv Perturb.	38.31	65.56	37.23	55.88	38.92	55.50	50.88	50.76	26.26	24.40
w/o Perturb.	46.32	69.08	41.25	65.05	57.66	68.64	50.88	50.66	26.23	24.38

Table 3: Ablation study results on four hallucination benchmarks.

Llama2-7B-Chat as the original model and fine-tune Llama2-7B-Base to obtain a factually weaker model. The hallucination model is trained on HaluEval (Li et al., 2023a). HaluEval dataset covers both factual and faithfulness hallucination types, and contains 35,000 hallucination samples across different task formats of fine-tuning data subsets, i.e., question answering (QA), summarization (Sum), dialogue (Dialog), and general instruction following (General), with 10,000 instances in QA, Sum, and Dialog and 5,000 in General. We use QA, Sum, Dialog, and their union (All) subsets for fine-tuning the hallucination model. We adopt LoRA (Hu et al., 2022) for parameter-efficient tuning and implement the pipeline with LLaMA-Factory (Zheng et al., 2024). For fine-tuning hallucination model, the perturbation radius ϵ is searched from $\{0.01, 0.1, 1\}$. We provide more detailed hyperparameter settings in Appendix A.

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 demonstrates the superiority of our model for hallucination mitigation. Specifically, for truthfulness-related datasets, compared to the baseline Greedy, RCD achieves improvements of **+12.4**%, **+19.8**%, and **+19.9**% on MC1, MC2, and MC3 scores on TruthfulQA. For knowledge-seeking tasks, RCD outperforms the baseline by **+4.7**% EM and **+4.4**% F1 scores on TriviaQA.

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 +2.01% improvement in *info*, a +6.57% 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 incorrect. This is because "reject" responses are often interpreted by the scoring model as fully correct, thereby receiving the maximum *truth* score. As a result, CD's overall *truth* score does not necessarily reflect factual accuracy.

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/o 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 four hallucination benchmarks 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 training data affect the method's mitigation performance. We evaluate four task formats corresponding to three HaluEval subsets, i.e., question answering (QA), summarization (Sum), dialogue (Dialog), and their combination (All). QA, Sum, and Dialog contains 10,000 samples, respectively. All aggregates the 30,000 samples from three subsets. We fine-tune the hallucination LLM on these data by using ICD and our RCD.

Table 4 shows results of ICD and our RCD against different task formats on four hallucination benchmarks. RCD outperforms ICD in most set-

Methods	Task Format	TruthfulQA MC1	TruthfulQA MC2	TruthfulQA MC3	FACTOR News	FACTOR Wiki	FACTOR Expert	TriviaQA EM	TriviaQA F1	NQ EM	NQ F1
RCD	Sum	**50.06**	**74.35**	**47.98**	64.96	56.52	66.10	50.69	50.44	25.90	24.06
	Dialog	49.69	70.24	44.05	**65.54**	59.07	69.91	**51.17**	**50.92**	26.43	24.54
	QA	43.08	70.19	43.02	65.44	**59.87**	**78.81**	51.12	50.87	**26.57**	**24.65**
	All	47.74	73.13	45.79	65.44	57.13	66.94	51.00	50.75	26.32	24.46
ICD	Sum	45.22	63.67	36.33	64.96	56.56	68.22	50.76	50.56	**26.23**	24.38
	Dialog	46.20	64.81	37.20	**65.05**	**57.66**	**68.64**	**50.88**	**50.66**	26.15	**24.44**
	QA	**46.32**	**69.08**	41.25	64.47	56.02	65.25	50.46	50.33	25.59	23.94
	All	41.73	67.74	**41.34**	64.48	56.26	65.87	50.78	50.56	25.96	24.03

Table 4: Results against different task formats of fine-tuning data on four benchmarks. We fine-tune the hallucination model on each subset of data using perturbation radius values of $\{0.01, 0.1, 1\}$, and report the best mitigation performance achieved on each benchmark.

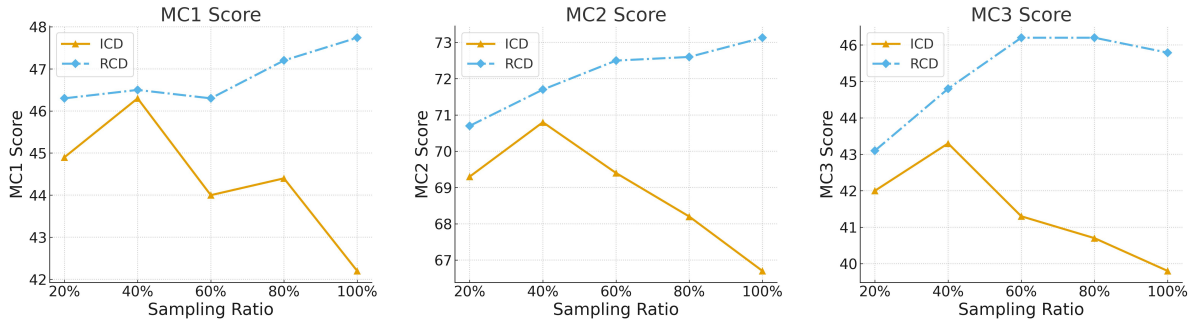

Figure 3: Results against different ratios of fine-tuning data for inducing hallucinations on TruthfulQA.

tings, demonstrating its effectiveness across different task formats for hallucination induction. RCD allows the weaker model to learn more diverse hallucination patterns across different task domains, achieving better hallucination mitigation.

Evaluation Across 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 samples) 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 of training samples, ICD struggles to learn sufficient hallucination patterns from limited data, leading to poor generalization. Our RCD can learn more diverse patterns from limited data by dynamically generating hard negative samples that better approximate the decision boundary of hallucinations. With a higher ratio of training samples, ICD tends to overfit to provide specific hallucination patterns for contrastive decoding, while RCD learns more generalized hallucination patterns, maintaining steadily improved mitigation performance.

Methods	TruthfulQA MC1	TruthfulQA MC2	TruthfulQA MC3
Baseline	37.62	54.60	28.12
ICD	46.32	69.08	41.25
RCD			
w/ FGM	**50.06**	**74.35**	**47.98**
w/ PGD	47.36	70.65	44.63

Table 5: Results against different attack methods for inducing hallucinations on TruthfulQA.

Evaluation against Different Perturbation Methods for Fine-tuning Factually Weak LLM 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 Method (FGM) and Projected Gradient Descent (PGD). FGM applies a single

Methods	TruthfulQA			FACTOR			TriviaQA		NQ	
	MC1	MC2	MC3	News	Wiki	Expert	EM	F1	EM	F1
Baseline	37.62	54.60	28.12	65.05	56.96	66.10	46.50	46.50	23.49	21.45
ICD	46.32	69.08	41.25	65.05	57.66	68.64	50.88	50.66	26.23	24.38
RCD										
$\epsilon = 0.01$	**50.06**	**74.35**	**47.98**	65.35	56.89	69.92	51.03	50.83	26.26	24.40
$\epsilon = 0.1$	48.10	70.78	45.82	65.35	57.10	71.61	50.99	50.76	26.37	24.43
$\epsilon = 1$	40.76	69.31	41.88	**65.44**	**59.87**	**78.81**	**51.17**	**50.92**	**26.57**	**24.65**

Table 6: Results against different perturbation radius when adversarially fine-tuning the hallucination model on four benchmarks. We adopt the subset of the optimal task formats on main results as fine-tuning data to build the hallucination model. For each benchmark, based on the same fine-tuning data, we experiment with different perturbation radius $\epsilon \in \{0.01, 0.1, 1\}$, and report the corresponding mitigation performance.

step perturbation along the normalized gradient direction, while PGD generates adversarial samples through iterative projected updates under a norm constraint. As shown in Table 5, RCD w/ FGM and w/ PGD consistently outperform comparison methods, highlighting the benefit of incorporating different adversarial perturbations in hallucination induction.

4.5 Parameter Analysis

We perform parameter analysis to study how the perturbation radius ϵ in adversarial fine-tuning affects the mitigation performance. The perturbation radius controls the magnitude of adversarial fine-tuning to generate worst-case samples when constructing the factually weaker model. We vary ϵ over $\{0.01, 0.1, 1\}$ during adversarial fine-tuning a factually weak LLM.

Experimental results on four benchmarks of our RCD against different ϵ are shown in Table 6. RCD achieves optimal mitigation results on TruthfulQA with $\epsilon = 0.01$, and $\epsilon = 1$ on other benchmarks. We also observe that, under different perturbation magnitudes, our RCD outperforms comparison methods on most evaluation metrics. This indicates that, by introducing hard negative samples, RCD learns more diverse hallucination features from the limited training data, leading to better mitigation results. We further provide parameter analysis of λ in Appendix B.2.

4.6 Effectiveness Evaluation Across Different LLM Scales

We evaluate the generalization capability of our proposed RCD method across large language models of varying sizes. Specifically, we compare the performance of the LLaMA2-7B model fine-tuned with 30K hallucination samples to larger LLaMA2 variants, including the 13B and 70B models.

Methods	TruthfulQA		
	MC1	MC2	MC3
Llama2-7B-Chat			
Baseline	37.62	54.60	28.12
ICD	46.32	69.08	41.25
RCD	**50.06**	**74.35**	**47.98**
Llama2-13B-Chat			
Baseline	37.75	55.67	28.16
ICD	48.47	73.47	46.04
RCD	**53.49**	**77.13**	**51.14**
Llama2-70B-Chat			
Baseline	37.70	58.99	29.79
ICD	51.04	75.01	46.54
RCD	**54.71**	**80.45**	**53.78**

Table 7: Performance comparison 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.

The results across different model sizes on TruthfulQA are shown in Table 7. RCD consistently outperforms the baseline across all model sizes, highlighting its scalability and strong generalization ability to larger language models.

4.7 Impact on Overall LLM 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., 2021) and ARC-Challenge (Clark et al., 2018). MMLU consists of multiple-choice questions covering a broad

Methods	MMLU	ARC-Challenge
Baseline	**0.472**	0.548
ICD	0.467	0.498
RCD	**0.472**	**0.550**

Table 8: Performance comparison of different decoding methods on overall LLM benchmarks.

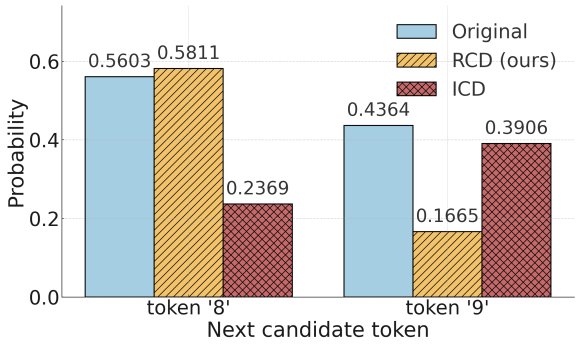

Figure 4: Token-level probability generated by the hallucination model for the query *"When was the rock and roll hall of fame built in Cleveland?"* from NQ dataset. The correct answer is *1995*, and a hallucinated answer is *1986*.

range of academic and professional subjects, testing general knowledge and factual reasoning. ARC-Challenge includes complex science questions that require multi-step reasoning. All experiments are conducted under the 5-shot setting to ensure consistency across methods.

Table 8 shows the results of different decoding methods on MMLU and ARC-Challenge. First, RCD outperforms ICD and achieves comparable performance to Baseline on MMLU. This demonstrates that our method does not compromise the model's general knowledge capabilities. Besides, RCD slightly outperforms the Baseline and ICD on ARC-Challenge, suggesting a potential benefit on complex question answering tasks.

4.8 Case Study

We provide a case study to illustrate how the contrast signal is formed and why RCD succeeds where ICD fails. Consider the query from NQ dataset: *"When was the rock and roll hall of fame built in Cleveland?"* The gold answer is *1995*, while a common hallucination is *1986*. In Figure 4, the bars labeled RCD and ICD report the weaker model's token probabilities for the two key next token candidates token '8' (from *1986*) and token '9' (from *1995*) which are contrasted against the original model's probabilities. Under RCD, the weaker model assigns relatively higher probability to the hallucination token '8' than the original model, and relatively lower probability to the correct token '9'; this yields a large positive contrast penalty for '8' and a small penalty for '9' in the distribution Equation (4), thereby suppresses the hallucinated '8' while preserving the correct '9'. In contrast, ICD makes the weaker model less confident on the correct token '9', which induces an excessive contrast penalty on '9' and mistakenly suppresses the true answer.

5 Conclusion

We present Regularized Contrastive Decoding (RCD), a new inference-time method that leverages hard negative samples to enhance contrastive decoding and achieve more effective hallucination mitigation. RCD learns diverse hallucination patterns to enhance the weaker model through adversarial-aware fine-tuning and employs contrastive decoding to mitigate hallucinations effectively. Experiments on four public hallucination benchmarks demonstrate that RCD consistently obtain better hallucination mitigation performance. Experiments also verified the effectiveness of RCD across different model sizes, task formats, perturbation methods and training data size.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No.U24A20335), the Fundamental Research Funds for the Central Universities (No.CUC25SG002), the China Postdoctoral Science Foundation (No.2024M753481), and Youth Innovation Promotion Association CAS. The authors thank the anonymous reviewers and the metareviewer for their helpful comments.

Limitations

Although our proposed RCD method effectively improves the factuality and reliability of LLM outputs, it requires additional computational resources for generating adversarial perturbations and fine-tuning a factually weaker model. Moreover, our evaluation focus on the Llama-2 family, and its effectiveness on other LLM architectures remains to be explored.

Ethical Considerations

Our method trains a factually weaker language model that is more prone to producing hallucinations. While this is effective for improving hallucination mitigation in LLMs, it raises ethical concerns: such a model could be misused to intentionally generate and spread misinformation or disinformation. To mitigate this risk, it should be handled responsibly and used only for research within controlled environments.

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.

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. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 7553–7567. PMLR.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Representation Learning*, volume 2024, pages 54158–54183.

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*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Dou Hu, Yinan Bao, Lingwei Wei, Wei Zhou, and Songlin Hu. 2023a. Supervised adversarial contrastive learning for emotion recognition in conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 10835–10852. Association for Computational Linguistics.

Dou Hu, Lingwei Wei, Yaxin Liu, Wei Zhou, and Songlin Hu. 2023b. UCAS-IIE-NLP at semeval-2023 task 12: Enhancing generalization of multilingual BERT for low-resource sentiment analysis. In *Proceedings of the The 17th International Workshop on Semantic Evaluation, SemEval@ACL 2023, Toronto, Canada, 13-14 July 2023*, pages 1849–1857. Association for Computational Linguistics.

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*.

Wentao Hu, Wengyu Zhang, Yiyang Jiang, Chen Jason Zhang, Xiaoyong Wei, and Li Qing. 2025. Removal of hallucination on hallucination: Debate-augmented RAG. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15839–15853, Vienna, Austria. Association for Computational Linguistics.

Xuming Hu, Junzhe Chen, Xiaochuan Li, Yufei Guo, Lijie Wen, Philip Yu, and Zhijiang Guo. 2024. Towards understanding factual knowledge of large language models. In *International Conference on Representation Learning*, volume 2024, pages 28680–28715.

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.

Xinyan Jiang, Hang Ye, Yongxin Zhu, Xiaoying Zheng, Zikang Chen, and Jun Gong. 2025. HICD: Hallucination-inducing via attention dispersion for contrastive decoding to mitigate hallucinations in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7764–7786, Vienna, Austria. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Jushi Kai, Tianhang Zhang, Hai Hu, and Zhouhan Lin. 2024. SH2: Self-highlighted hesitation helps you decode more truthfully. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4514–4530, Miami, Florida, USA. Association for Computational Linguistics.

Diederik Kinga, Jimmy Ba Adam, et al. 2015. A method for stochastic optimization. In *International conference on learning representations (ICLR)*, volume 5. California;.

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, 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, Singapore. Association for Computational Linguistics.

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.

Takeru Miyato, Andrew M. Dai, and Ian J. Goodfellow. 2017. Adversarial training methods for semi-supervised text classification. In *ICLR (Poster)*.

Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov, Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. 2024. Generating benchmarks for factuality evaluation of language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 49–66, St. Julian's, Malta. Association for Computational Linguistics.

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

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. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7744–7757, Miami, Florida, USA. Association for Computational Linguistics.

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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand. Association for Computational Linguistics.

A Details of Hyperparameter Settings

We train Llama2-7B-Base with Adam optimizer (Kinga et al., 2015), a learning rate of 5×10^{-4} with zero warmup ratio, a total batch size of 256 samples, and enable LoRA adapters on attention projections (`q_proj`, `k_proj`, `v_proj`). We summarize the hyperparameter settings on each benchmark in Table 9.

Configuration	TruthfulQA	FACTOR	TriviaQA	NQ
Number of epochs	5	5	5	5
Total batch size	256	256	256	256
Optimizer	Adam	Adam	Adam	Adam
Learning rate	$5e^{-4}$	$5e^{-4}$	$5e^{-4}$	$5e^{-4}$
Warmup ratio	0.0	0.0	0.0	0.0
Perturbation radius	0.1	1	1	1
Fine-tuning data	Sum	QA	QA	Dialog

Table 9: Hyperparameter settings of RCD on four benchmarks.

B Supplementary Experimental Results

B.1 Efficiency Analysis

We compare the inference efficiency of different inference-time 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 fine-tuned Llama2-7B-Base weaker model.

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

Table 10: Inference time comparison across different decoding strategies.

Table 10 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 fine-tuned 7B-Base weaker model, the overhead is roughly 2.91× and 2.78× respectively. Although these configurations

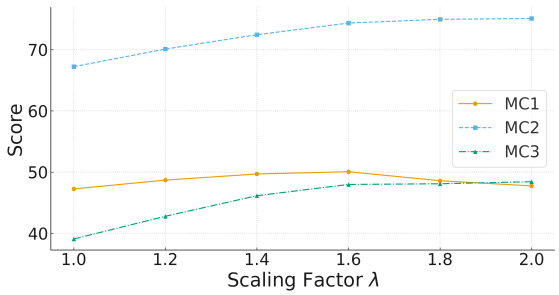

Figure 5: MC1, MC2, and MC3 scores on the TruthfulQA dataset for different scaling factors λ.

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.

B.2 Supplementary Parameter Analysis

To better understand the behavior of RCD, we analyze the scaling factor λ, which controls the strength of the contrastive signal from the weaker model. Unless otherwise noted, this analysis uses the TruthfulQA benchmark with the weak model adversarially fine-tuned on the HaluEval *summarization* subset and an L_2-normalized FGM magnitude $\epsilon = 0.01$. Figure 5 plots MC1/MC2/MC3 as λ varies from 1.0 to 2.0. Increasing λ amplifies the penalty from the weaker (hallucination) model, thereby strengthening hallucination suppression and improving accuracy up to a point. Empirically, $\lambda = 1.6$ offers a strong trade-off across metrics (MC1 50.06, MC2 74.35, MC3 47.98): larger values can slightly boost MC2/MC3 but start to reduce MC1, indicating over-penalization that suppresses some potentially correct tokens.