Harder is Better: Hard Hallucination-Induced Contrastive Decoding for Hallucination Mitigation

Anonymous ACL submission

Abstract

Large Language Models (LLMs) 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 HiCD method to improve hallucination mitigation. It aims to inject hard-to-detect hallucinations to enhance the robustness of contrastive decoding during inference. An adversarial-aware strategy is introduced for finetuning hallucination models to effectively learn more precise and diverse hallucination patterns from available hallucination data. This enhances the contrastive decoding process, enabling more effective identification and filtering of erroneous content. We evaluate HiCD on four various hallucination benchmarks. Experimental results show significant improvements on all metrics consistently, proving the effectiveness and superiority of HiCD for hallucination mitigation.

1 Introduction

Large Language Models (LLMs) have demonstrated substantial progress in a 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). Such hallucinations pose significant risks, particularly in high-stakes domains such as legal consultation, medical advice, and specialized technical support, where factual reliability is essential.

Various strategies have been pursued to mitigate hallucinations. Some works emphasized data-

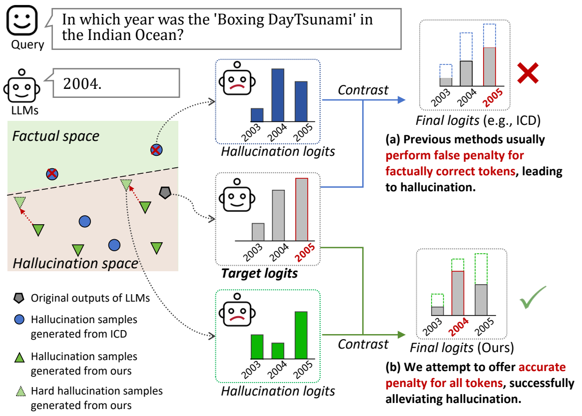

Figure 1: An illustration showing how over-penalization of factually correct tokens leads to hallucination

centric methods, such as curating training sets or integrating external knowledge to guide models toward factual correctness (Sun et al., 2023; Shuster et al., 2021). These methods typically require substantial computational overhead and may not generalize well beyond the data distributions observed during training (Ren et al., 2023; Borgeaud et al., 2022). Recently, increasing attention has focused on mitigating hallucinations at the *inference* stage (Li et al., 2023b; Chuang et al., 2023; Kai et al., 2024; Zhang et al., 2023). They usually examine differences across multiple candidate outputs via contrastive decoding strategies for hallucination mitigation during inference. Inference-stage methods can be more flexible and less resource-intensive than strategies that rely solely on enhancing training data or model parameters.

However, the above inference-stage methods may suffer from the precision of hallucination tokens, leading to limited contrastive performance during inference. Specifically, hallucinations in LLMs are highly diverse (Huang et al., 2023). Finetuning with the scarcity of hallucination data often leads to a suboptimal hallucination model, which struggles to generalize well and fails to provide subtle and precise hallucination patterns (Wang et al., 2023). As a result, factually accurate tokens are prone to be over-penalized during contrastive decoding, leading to suboptimal performance for hallucination mitigation. As shown in Figure 1, the imprecise hallucination logits outputted by previous works may perform false penalty for the factual-correct token (i.e., *2024*). Therefore, a more effective fine-tuning strategy for hallucination model needs to be explored for capturing more precise and diverse hallucination samples, accordingly to improve the effectiveness of contrastive decoding.

In this paper, we propose a new inference-stage method, Hard Hallucination-Induced Contrastive Decoding (HiCD), to improve hallucination mitigation. Our HiCD aims to inject *hard*-to-detect hallucinations to enhance the robustness of contrastive decoding during inference. We design a new adversarial-aware finetuning strategy for hallucination models to explore more hard hallucination samples. These samples usually are similar to factually correct tokens but deviate in subtle ways. As shown in Figure 1, they lie near the decision boundary between factual correctness and hallucination in the model's prediction space. To achieve this, we utilize adversarial perturbations to encourage factually correct samples beyond the limited hallucination dataset to more accurately approach hallucination boundaries (Goodfellow et al., 2014). This process reduces the prediction probabilities of correct tokens in a controlled manner, preventing the model from overfitting to specific hallucination patterns. Based on hallucination LLMs finetuned by our strategy, during the contrastive decoding phase, the model avoids erroneously penalizing factually correct tokens, resulting in outputs that are more reliable and factually consistent. Importantly, HiCD achieves these improvements without requiring extensive data curation or large-scale retraining, offering a scalable and practical solution for mitigating hallucination issues in LLMs.

We conduct experiments on four truthfulness assessments and knowledge-seeking datasets for hallucination alleviation evaluation. The experimental results demonstrate HiCD's effectiveness, with consistent improvements on multiple benchmarks (e.g., +4.08% MC2 on TruthfulQA and +9.03% on FACTOR Expert) across diverse tasks. Additionally, ablation and parameter analyses highlight the crucial role of adversarial training and optimal hyperparameters, indicating HiCD's broad applicability for enhancing factual fidelity and mitigating hallucinations in large language models.

Our contributions are threefold: 1) we propose a new inference-stage HiCD method to improve hallucination mitigation. It injects hard-to-detect hallucinations 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 HiCD 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.

Initial efforts to mitigate hallucinations often emphasized data- and model-centric strategies. Data-centric approaches involve refining training corpora—either curating higher-quality data or incorporating external knowledge sources—to encourage factual correctness (Sun et al., 2023; Shuster et al., 2021; Lin et al., 2022). Model-centric methods aim to modify training objectives, sometimes leveraging techniques like reinforcement learning from human feedback to align model outputs with human judgment (Wang and Sennrich, 2020; Ouyang et al., 2022). While these methods can reduce certain types of hallucinations, they often require extensive data preparation, large-scale retraining, may not generalize well to complex, subtle errors that lie near decision boundaries.

To address these issues more efficiently, researchers have turned to inference-stage interventions. Post-hoc decoding strategies can be applied at generation time without modifying the underlying parameters. By using contrastive signals or

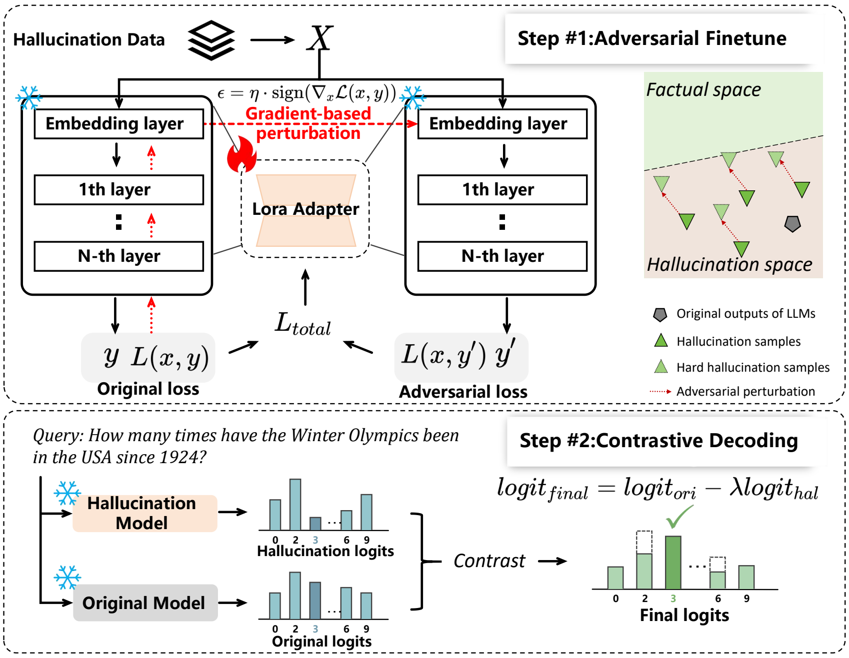

Figure 2: Overview of our HiCD framework. In the adversarial finetuning phase, we induce hard-to-detect hallucinations through gradient-based perturbations, resulting in a weaker "hallucination" model. During inference, contrastive decoding combines outputs from the original and hallucination models,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

other dynamic generation criteria, these approaches aim to identify and filter out hallucinatory content as it emerges (Chang et al., 2023). However, existing inference-stage methods often rely on hallucinations that are easy to induce or naturally occurring. Such limited sets of negative examples fail to represent the full spectrum of challenging hallucinations that can occur in practice. As a result, these methods struggle with difficult, subtle hallucinations in more complex, real-world scenarios.

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. (2023) 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, such as Induce-then-Contrast Decoding (ICD), are constrained by the limited availability of hallucination data, which is insufficient to fully influence the extensive knowledge acquired by large language models during pretraining. This limitation hampers their ability to effectively identify and mitigate subtle or complex hallucinations that closely resemble truthful content. Consequently, these methods may inadvertently penalize factually correct tokens, reducing their accuracy and reliability in real-world applications where distinguishing between factual information and fabrications is critical. Addressing these challenges requires more sophisticated strategies that can generate richer and more nuanced negative examples, thereby enabling a more precise approximation of the true decision boundaries between accurate and erroneous outputs.

3 Hard Hallucination-Induced Contrastive Decoding (HiCD)

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, Hard Hallucination-Induced Contrastive Decoding (HiCD), aims to reduce hallucinations by leveraging contrastive decoding between a strong model and a weaker, adversarially trained model.

3.1 Inducing Hard Hallucinations

As hallucinations in LLMs are highly diverse and subtle. Previous works (Zhang et al., 2023) inducing potential hallucination in a suboptimal way usually falsely penalized precision factual tokens, leading limited alleviation performance. To capture hard hallucination samples for better contrasting, we design a new adversarial-aware finetuning strategy to capture hard hallucination samples for better contrastive decoding during generation of target LLMs. Specifically, we first employ few-shot prompting techniques to generate misleading or incorrect responses from a factual dataset. We then go further by integrating adversarial training to push the weaker model—referred to as the "hallucination LLM"—towards producing more intricate, boundary-like hallucinations that are harder to distinguish from truthful outputs.

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. The initial fine-tuning objective is:

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

where θ denotes the original model parameters. After this step, we incorporate adversarial perturbations to shape $\Delta\theta$ so that the weaker model becomes more inclined to produce complex hallucinations.

By introducing adversarial training during fine-tuning, the weaker model's errors become more refined and deceptive, rather than simple and easily detectable. We employ the Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2015) to perturb the input embeddings $\mathbf{x}$:

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

where ϵ controls the perturbation magnitude, and $\mathcal{L}$ is the loss function. This pushes the model's decision boundaries, increasing uncertainty and promoting the production of subtle hallucinations.

Training alternates between clean and adversarially perturbed examples. The combined objective is:

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

resulting in a hallucination LLM that naturally generates a richer, set of more challenging negative examples for the subsequent contrastive decoding step.

3.2 Contrastive Decoding

Having obtained the stronger model θ and the adversarially fine-tuned weaker model θ', we apply contrastive decoding (Li et al., 2023b) to their outputs. At each timestep t, both models compute the conditional probability of the next token x_t. We define the contrastive score as:

$$\mathcal{F}_t = \log p(x_t \mid x_{<t}; \theta) - \lambda \log p(x_t \mid x_{<t}; \theta'), \qquad (4)$$

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, hallucination-prone Evil 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:

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

where $\gamma \in (0, 1]$ is a hyperparameter that determines the filtering threshold. This ensures that only tokens whose log probabilities are within $\log \gamma$ of the highest log probability are retained.

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

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

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.

Setting	Value
Model	Llama2-7B-Base
Epochs	5
Device	NVIDIA Tesla A100 80GB
Total Batchsize	256
Learning Rate	5×10^{-4}
LoRA Target	$q_{\text{proj}}, k_{\text{proj}}, v_{\text{proj}}$

Table 1: Finetuning settings for building the factually weaker model.

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., 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.

Comparison Methods. We evaluate the effectiveness of our proposed method by comparing it against the following baselines: (1) **Greedy Decoding**: A default approach where the highest probability token is selected at each step without additional decoding techniques. (2) **Induced Task Inference (ITI)** (Li et al., 2024): This method enhances generalization by applying task-specific adjustments during inference, refining predictions based on task-relevant cues. (3) **Contrastive Decoding (CD)** (Li et al., 2023b): Aims to reduce hallucinations by contrasting outputs from a strong model and a weaker model, emphasizing reliable predictions while penalizing non-factual ones. (4) **Direct Output Layer Adaptation (DoLa)** (Chuang et al., 2023): Focuses on adjusting the model's output layer to improve factual accuracy, particularly for knowledge-intensive tasks. (5) **Induce-then-Contrast Decoding (ICD)** (Zhang et al., 2023): Integrates hallucination induction with contrastive decoding, leveraging a weakened model to penalize incorrect predictions and reinforce factual outputs. (6) **Activation Decoding (AD)** (Shi et al., 2024): Amplifies the influence of contextual information over a language model's prior knowledge by employing a contrastive output distribution, improving faithfulness in tasks requiring external knowledge integration.

Implementation Details All experiments are conducted on a single NVIDIA Tesla A100 80GB GPU using the Llama2 series models. The scaling factor λ in Equation 4 was set to 1.8 for optimal results on the TruthfulQA dataset. For the FACTOR dataset, the best results were achieved with λ values of 0.35. 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., 2023). Specifically, we use the HaluEval dataset(Li et al., 2023a) to fine-tune the weaker model. LoRA (Hu et al., 2022) is used for parameter-efficient fine-tuning and hallucination injection. The LLaMA-Factory framework (Zheng et al., 2024) is also employed for fine-tuning. Details of the fine-tuning process and hyperparameters are provided in Table 1.

4.2 Main Results

Overall results on four datasets for hallucination mitigation are shown in Table 2. The proposed HiCD achieves the best performance on all datasets

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., 2023)	46.32	69.08	41.25	70.75	58.40	66.94	50.46	50.33	25.59	23.94
HiCD (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 2: 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 HiCD, we follow Zhang et al. (2023) and finetune Llama2-7B-Base as a weaker model for contrasting with Llama2-7B-Chat. The best performances are **bolded**.

in terms of all evaluation metrics. This demonstrate the superiority of our model on ensuring the truthfulness of responses but also effectively retrieving and reasoning over factual information in open-domain settings. Specifically, for truthfulness-related datasets, compared the the baseline Greedy, HiCD achieves improvements of **+9.4%**, **+18.6%**, and **18.1%** on MC1, MC2, and MC3 scores on TruthfulQA. For knowledge-seeking tasks, HiCD outperforms the baseline by **+4.4%** EM and **4.2%** F1 scores. Besides, compared to other decoding strategies, HiCD contrasts with hard hallucination-induced models, leading to better mitigation performance on all datasets.

4.3 Ablation Study

To evaluate the effectiveness of our adversarial training in inducing precise hallucinations and enhancing contrastive decoding, we conduct an ablation study by comparing our HiCD with the following ablation models: 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 HiCD 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, HiCD enables more effective filtering of factual inaccuracies, leading to more reliable and factually consistent outputs.

Method	TruthfulQA			FACTOR		
	MC1	MC2	MC3	News	Wiki	Expert
HiCD	**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 results on TruthfulQA and FACTOR.

Method	TruthfulQA			
	%truth	%info	%truth*info	%reject
CD	70.21	42.25	19.23	29.98
ICD	62.85	77.65	41.16	23.50
HiCD	63.71	78.03	42.24	23.13

Table 4: Evaluation results on generative tasks using "GPT-judge" for TruthfulQA.

4.4 Generation Task Evaluation

Following Lin et al. (2022), we also evaluate our method on the TruthfulQA dataset using "GPT-judge" to assess both factual accuracy and informativeness. This evaluation yields four metrics: *truth*, *info*, a combined *truth&info*, and the *reject* rate. Table 4 presents the evaluation results on generative tasks for CD, ICD, and our proposed HiCD approach. Compared to ICD, HiCD achieves a +0.38% increase in *info*, a +1.08% increase in *truth&info*, and a -0.37% decrease in *reject*, indicating that HiCD produces more informative, factually consistent responses.

4.5 Efficiency Analysis

We compare the inference efficiency of different inference-stage methods, i.e., a baseline greedy decoding, CD, ICD, and our proposed HiCD. 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 the CD experiment uses a Llama2-13B-Chat vs. 7B-Chat configuration, while both ICD and HiCD

Method	Decoding Latency (s)
Baseline	138.4 ($\times$1.00)
CD	357.6 ($\times$2.58)
ICD	402.4 ($\times$2.91)
HiCD	384.7 ($\times$2.78)

Table 5: Inference time comparison across different decoding strategies.

rely on a Llama2-7B-Chat model with a finetuned Llama2-7B-Base weaker model.

As shown in Table 5, the baseline decoding takes approximately 138.4s. Under the CD setting, increasing complexity leads to about a 2.58× slowdown. For ICD and HiCD, 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, HiCD offers improved factual fidelity over ICD while slightly reducing the slowdown from the baseline, indicating a more balanced trade-off between accuracy and efficiency.

4.6 Parameter Analysis

We experiment to analyze the impact of two critical hyperparameters in HiCD: the perturbation magnitude ϵ and the scaling factor λ. Results of parameter analysis on TruthfulQA are shown in Figure 3.

Effect of Scaling Factor λ 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 1.5. 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.

Effect of Perturbation Magnitude ϵ The perturbation magnitude ϵ controls the strength of adversarial noise during the fine-tuning of the weaker model. By adjusting ϵ, we influence the extent to which the model's decision boundaries are shifted, thereby affecting the precision and difficulty of induced hallucinations. Our results indicate that

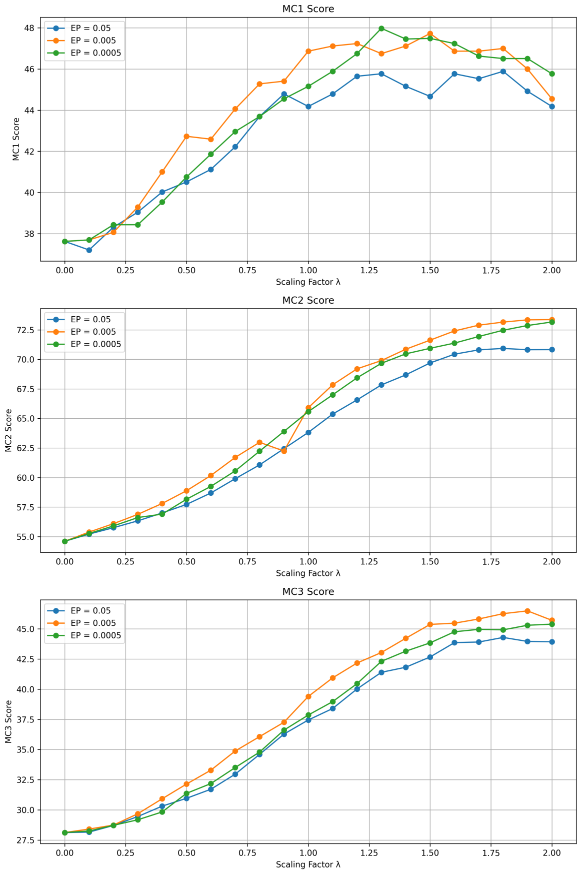

Figure 3: MC1, MC2, and MC3 scores on the TruthfulQA dataset for different perturbation magnitudes ϵ and scaling factors λ.

$\epsilon = 0.005$ yields the highest MC scores, effectively balancing the generation of challenging hallucinations and maintaining the efficacy of contrastive signals. Smaller perturbations ($\epsilon = 0.0005$) do not sufficiently alter the model's behavior to produce hard hallucinations, while larger perturbations ($\epsilon = 0.05$) may overly degrade the weaker model's performance, reducing the effectiveness of contrastive decoding in distinguishing factual from hallucinated content.

4.7 Case Study

We provide a case study from the Natural Questions dataset to illustrate the effectiveness of our method. Consider the query: *"When was the rock and roll hall of fame built in Cleveland?"* The correct answer is *1995*, while a hallucinated answer is *1986*. In this scenario, both the original model and the ICD approach produce the hallucinated answer, whereas our method yields the factually correct output. As shown in Figure 4, we analyze the token-level probabilities for the key differing token positions (the second "9" in *1995* and "8" in

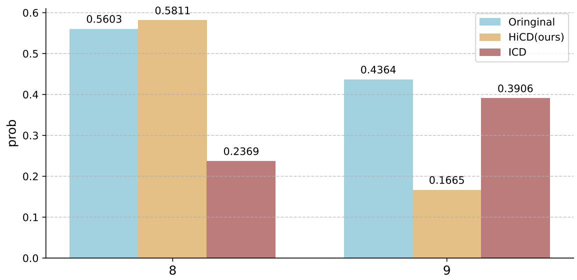

Figure 4: Token-level probability analysis for the query *"When was the rock and roll hall of fame built in Cleveland?"* at the critical token position where hallucination occurs

1986): the original model assigns overly high confidence to an incorrect token, while ICD's weaker model overcompensates for the correct token, ultimately leading to a hallucination. In contrast, our weaker model appropriately balances probabilities for the correct and hallucinated tokens, ensuring that the final output is both accurate and reliable.

5 Conclusion

We presented Hard Hallucination-Induced Contrastive Decoding (HiCD), a novel inference-stage method that leverages adversarial perturbations to induce more challenging hallucinations for improved contrastive filtering. By doing so, HiCD significantly enhances factual fidelity and robustness across multiple benchmarks, including TruthfulQA, FACTOR, TriviaQA, and NQ. More precise and diverse signals are produced by HiCD consistently outperform state-of-the-art baselines, offering a scalable and practical approach to mitigating hallucinations in large language models.

6 Limitations

While our proposed HiCD 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.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

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*.

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.

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*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.

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.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2023. Investigating the factual knowledge boundary of large language models with retrieval augmentation. *arXiv preprint arXiv:2307.11019*.

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.

Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Haojun Huang, Wei Ye, Xiubo Geng, et al. 2023. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. *arXiv preprint arXiv:2302.12095*.

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*.

Yue Zhang, Leyang Cui, Wei Bi, and Shuming Shi. 2023. Alleviating hallucinations of large language models through induced hallucinations. *arXiv preprint arXiv:2312.15710*.

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*.