Attributive Reasoning for Hallucination Diagnosis of Large Language Models

Yuyan Chen[1], Zehao Li[2], Shuangjie You[3], Zhengyu Chen[4], Jingwen Chang[1], Yi Zhang[5], Weinan Dai[4], Qingpei Guo[6], Yanghua Xiao[1]*

[1]Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University
[2]School of Data Science and Engineering, East China Normal University
[3]Georgia Institute of Technology
[4]Zhejiang University
[5]Southern University of Science and Technology
[6]Ant Group
{chenyuyan21@m., jwchang24@m., shawyh@}fudan.edu.cn, {leepaul.private,wnd17460,youshuangjie}@gmail.com,
zhangyi2021@mail.sustech.edu.cn, chenzhengyu@zju.edu.cn, qingpei.gqp@antgroup.com

Abstract

In recent years, large language models (LLMs) have demonstrated outstanding capabilities in various tasks. However, LLMs also have various drawbacks, especially hallucination. Hallucination refers to the generation of content that does not align with the user input, contradicts previously generated content or world knowledge. Current research on hallucination mainly include knowledge retrieval, prompt engineering, training data improvement, reinforcement learning, etc. However, these methods do not involve different categories of hallucinations which is important on hallucination analysis, and make detailed investigation for the internal state of LLMs which indicates the direction on hallucination occurrence. Therefore, in our research, we introduce an attribution framework to trace the origins of hallucinations based on the internal signals of LLMs. To support this framework, we develop a new benchmark named RelQA-Cate, which includes eight categories of hallucinations for the answers generated by LLMs. After that, we present a novel Differential Penalty Decoding (DPD) strategy for reducing hallucinations through adjusting post-probabilities of each answer. We conduct a series of experiments and the performance on answer reliability has significant improvement, achieving 28.25% at most, which demonstrates the effectiveness of our proposed DPD and its generalization in mitigating hallucination in LLMs.

Introduction

Large language models (LLMs) have a wide range of applications in various downstream tasks (Chen et al. 2024c,g,d,a). However, it is important to recognize that while LLMs bring about revolutionary technological advances, they also bring a series of security and privacy issues that deserve our attention, especially hallucinations (Ji et al. 2023). Hallucinations refer to the circumstances where the content generated by LLMs is incorrect to the given question, unrelated or even conflicting to the input prompt. They have various categories (Huang et al. 2023), such as factual error, representing the generated text contains obvious factual errors. For instance, in response to the question, "Who was the first president of the United States?", an incorrect answer is "Aaron Burr". Another category is conceptual confusion which indicates LLMs mix up different concepts in responses, such as considering the solar cell as a type of "electronic device" instead of "electrical device" [1]. There are also more categories of hallucinations which are not fully investigated and the underlying reasons for each category related to LLMs' internal state are even largely uncharted.

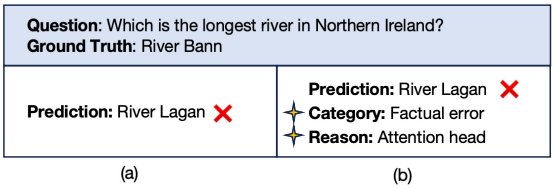

Figure 1: The predicted answer without and with hallucination category for a given question.

Understanding the reasons behind each category of hallucination in LLMs is crucial. Hallucination categories can be likened to *symptoms* in medicine, while the reasons of these hallucinations can be compared to the *etiology* in medical diagnoses. In the case of hallucinations, these causes are typically linked to LLMs' internal states (i.e. representations of the input). For instance, as shown in Fig. 1, if we not only identify hallucination occurs based on wrong prediction "River Lagan" (see Fig. 1 (a)), but also recognize this is a "Factual error," which is possibly due to the LLMs overly focusing on a certain attention head of a particular layer (see Fig. 1 (b)), we can address the issue by adjusting the attention distribution of that layer or retraining the relevant head.

However, current research mainly focus on other methods except investigating the internal state of LLMs in reducing hallucinations, including knowledge retrieval (Chen, Xiao, and Liu 2022; Shi et al. 2023; Peng et al. 2023; Chen et al. 2024b,f), prompt engineering (Zhang et al. 2023; Tou-

*Corresponding author.

[1]Solar cell is more commonly considered electrical devices, as its primary function is the generation and supply of electrical power. Electronic devices generally contain computers, mobile phones, and radio receivers, etc.

vron et al. 2023; Chen et al. 2023f), training data improvement (Zhong et al. 2021; Chen et al. 2023b; Cao, Kang, and Sun 2023; Chen et al. 2023g), reinforcement learning (Yu et al. 2023; Sun et al. 2023; Chen et al. 2024e). But knowledge retrieval may not fully encompass all areas or the latest information. Prompt engineering has high demands on how users input their queries. Improving training data performs bad when faced with novel situations. Reinforcement learning may need a complex training process. A better approach is to examine the differences in the internal states of LLMs when producing correct answers and hallucinations. Adjustments can be made to the LLM's output to specifically target and reduce hallucinations.

Therefore, we propose an attribution framework that traces the reasons of hallucinations generated by LLMs based on their internal signals. The internal causes in our research represent the differences between hallucination output and the correct output based on the representation of hidden layers, self-attention outputs, and high-contribution words, among others. To support this framework, we first design eight hallucination categories inspired by Li et al. (2024) and classify the incorrect answers generated by LLMs into these categories using ChatGPT based on the RelQA dataset (Chen et al. 2023d) to obtain a new benchmark named RelQA-Cate. Under the guidance of the framework and the benchmark, we realize a novel strategy named Differential Penalty Decoding (denoted as DPD) [2]. The significance of DPD is that it does not require additional annotation costs and helps open the black box of LLMs. We conduct extensive experiments, demonstrating that DPD has a great effect across various datasets and LLMs in mitigating hallucinations.

Attribution Framework

To trace the causes of LLMs producing hallucinations, we annotate the typical categories of hallucinations and reveal the internal states of LLMs to establish a hallucinations attribution framework.

Hallucination Category Annotation

We adopt several powerful LLMs, including LLaMA-7B, LLaMA2-7B, Baichuan-7B and Mistrial-7B to generate answers in RelQA (Chen et al. 2023d) dataset. We adjust the temperature as 0 to generate unchangeable answers. Next we adopt ChatGPT [3] to categorize hallucinations of answers generated by LLMs into eight categories including *Factual Error Hallucination*, which represents a text includes incorrect facts; *Logical Error Hallucination*, which represents an answer is illogical or contradicts itself; *Conceptual Confusion Hallucination*, which represents an answer mixes up different concepts; *Vagueness Hallucination*, which represents an answer that is overly vague; *Lack of Commonsense Hallucination*, which represents an answer contradicts commonsense; *Over-generalization Hallucination*, which represents an answer is too broad and lack detail; *Emotional Bias Hallucination*, which represents an answer exhibits prejudice or an emotional tone; *Lack of Uncertainty Hallucination*, which represents an answer is overly confident about uncertain events. Because some hallucination types might overlap, we focus on the most significant ones. The samples are shown in Table 1.

In this process, we conduct categorizing validation with three human evaluators. We have human evaluators inspect the categorization; if there is disagreement, we revise the categorization. For inconsistent opinions, we adopt the majority categorization result. If each evaluator produces different results, the data point is discarded. Finally, we select 1,500 data instances for each hallucination category in RelQA, ensuring an equal distribution of 12,000 correct answers, thereby constructing a dataset of 24,000 samples, named RelQA-Cate, serving as an evaluation dataset for assessing LLMs' hallucinations attribution.

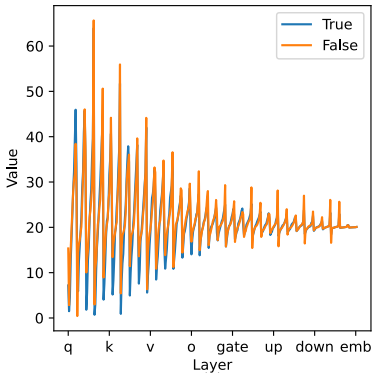

Figure 2: Visualization of hidden layer output of the first category of hallucination. The horizontal axis represents different layers of the LLM's internal structure, and the vertical axis represents the absolute output values of those layers.

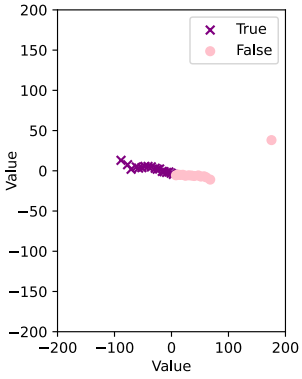

Figure 3: Visualization of two-dimensional hidden layer output of the first category of hallucination. The horizontal and vertical axes represent the absolute output values after feature dimensionality reduction.

[2]https://github.com/Yukyin/DPD4LLM

[3]https://chat.openai.com

Category	Question	Ground-truth	Predicted answer with hallucination
Factual error	Which is the longest river in Northern Ireland?	River Bann	River Lagan
Logical error	Who was the director of the 2001 French action horror film starring Mark Dacoscos?	Christophe gans	Mark Dacascos is an American actor and martial artist.
Conceptual confusion	Which is the third largest of the world's oceanic divisions, covering 70,560,000 km2 (approximately 20% of the water on the Earth's surface), and bounded by Asia on the north, on the west by Africa, on the east by Australia, and on the south by the Southern Ocean or, depending on definition, by Antarctica?	Indian ocean	The Pacific Ocean is the largest and deepest of Earth's oceanic divisions. It extends from the Arctic Ocean.
Vagueness	At the same time, Cho wrote and published an autobiographical book with the same title, and the show itself was filmed and released as a concert film in 2000.... What happened in 2000?	Cho wrote and published autobiographical book with same title and show itself was filmed and released as concert film in 2000	im one that i want was filmed and released as concert film in 2000.
Lack of commonsense	What do birds typically build to lay their eggs?	Nests	Tunnels
Over-generalization	Both Bessera and Sesleria are?	genus of perennial plants	plants
Emotional bias	Al-Askari's mother, as in the case of the majority of The Twelve Imams, was a slave girl who was honoured after bearing children with the title Umm walad (mother of offspring). ... Why did the slave girl cry	can not answer	alaskaris mother slave girl cried because she had seen her son being endowed with divine knowledge which led him to ponder deeply about purpose of life and afterlife
Lack of uncertainty	Nowsch, 19, faces three felony charges: murder, attempted murder and unlawful discharge of a gun from a vehicle. How did he kill?	shooting	I'm not sure.

Table 1: Samples with different categories of hallucination.

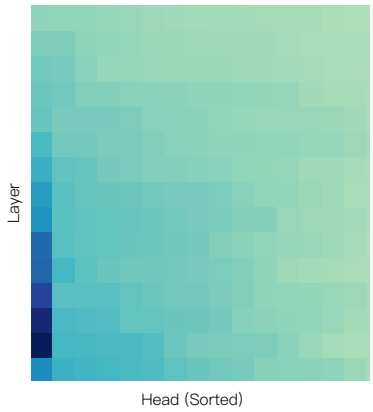

Figure 4: Visualization of attention output of the first category of hallucination. The horizontal axis represents sorted LLM's attention heads, and the vertical axis represents the indices of the attention layers.

Internal State Analysis

We choose LLaMA2-7B (Touvron et al. 2023) as our backbone and RelQA as an example to visually analyze the differences between each category of hallucination output and the correct output.

Hidden layer outputs. We first analyze the absolute values of each hidden layer's output across each category of hallucinations to assess the sensitivity of different hidden layers to different categories. Specifically, for a given question, we record outputs of each hidden layer of an LLM while it produces each category of hallucination in the predicted answer. We average results of a large number of samples for analysis, expecting to identify which hidden layers are more active when dealing with specific hallucinations. As shown in Fig. 2, the blue line represents correct answers and the orange line represents incorrect answers with factual error hallucination. We then reduce the high-dimensional outputs of the hidden layers to two-dimensional spaces using PCA for a coarse-grained analysis, as depicted in Fig. 3 representing factual error hallucination. We observe that the purple points (i.e. correct answers) and pink points (i.e. incorrect answers with factual error hallucination) are clearly separated, indicating considerable contribution of hidden layers for model generating correct answers.

Self-attention outputs. Next, we analyze the self-attention output of each hidden layer and head for each category of hallucination to identify differences in the LLM's focus. As shown in Fig. 4, we observe that in the higher layers, attention heads show specialized focus, with some heads concentrating on specific regions, indicating an emphasis on global information and higher-level feature integration. The attention patterns suggest that the need for refining attention mechanisms to reduce inaccuracies in generated content.

After that, we calculate the correlation between normalized self-attention outputs of each category of hallucination with Pearson correlation coefficient. As shown in Fig. 5, the coefficients between various categories of hallucinations and correct answers are generally high, indicating that the LLMs often struggles to differentiate between incorrect and correct answers. High correlation values may be one of the factors contributing to the occurrence of hallucinations.

High-contribution words. Finally, we identify high-contribution words of the each category of hallucination. We adopt gradient-based approaches (Simonyan, Vedaldi, and Zisserman 2013) to compute saliency map based on the gra-

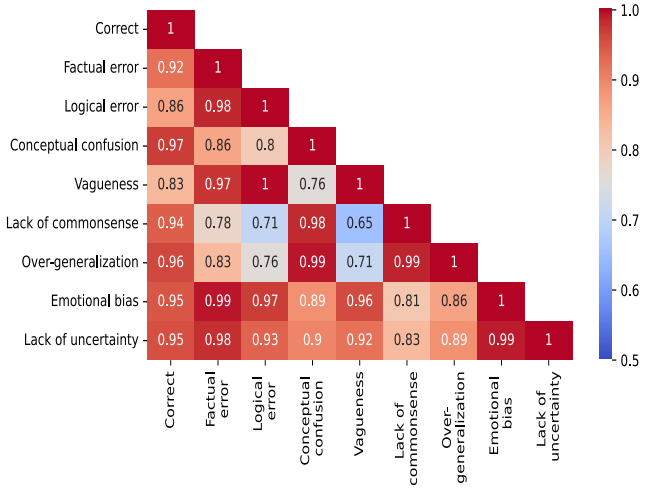

Figure 5: Correlation of attention output of each category of hallucination.

Factual error	#s Which is the longest river in Northern Ireland ?

Figure 6: Salient words of the first category of hallucination.

dient of the input questions with respect to the generated answers. After that, we use a visualizer to present saliency maps for saliency scores to find high-contribution words. For example, for the first category as shown in Fig. 6, we observe "longest river" and "Northern Ireland" are emphasized. It could be an explanation of why the LLM produces hallucinations.

Methodology

In this section, we propose a novel decoding strategy named Differential Penalty Decoding (denoted as DPD) as shown in Fig. 7.

Generating candidate answers. We first adopt an LLM to generate k (set as 5) diverse candidate answers through adjusting the temperature coefficient. The diversity degree of candidate answers for a question is evaluated by Distinct-2 (Li et al. 2015), a metric for assessing text diversity by calculating the proportion of unique bi-grams in generated text, which is requested over α (set as 0.8). These candidates reflect the LLM's internal state.

Calculating penalty values. Next, we calculate the corresponding penalty values for each candidate answer based on LLM's internal state through above-mentioned five dimension. We start by analyzing the absolute values of the outputs of hidden layers. To adjust the outputs of these layers to be closer to the activity levels seen when processing correct answers, we design the following penalty function:

$$P_l = \sum_{j=1}^{N} \sum_{i=1}^{L} (o_{ij} - \bar{o_i}) \cdot \text{sign}(o_{ij} - \bar{o_i}), \tag{1}$$

where P_l represents the penalty term, o_{ij} represents the absolute value of the output from the i-th layer when processing the j-th category of hallucination, $\bar{o_i}$ signifies the average absolute output value of that same layer when generating correct answers, L is the total number of hidden layers in the LLM, N is the number of hallucination categories. Sign function is to ensure that when the output value is above the mean, the penalty is positive, and when it is below the mean, the penalty is negative.

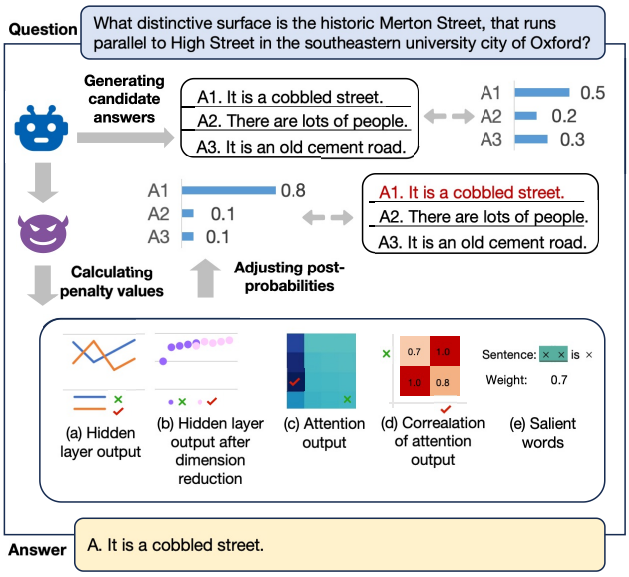

Figure 7: The decoding strategy named DPD based on the attribution framework.

Next, through PCA dimensionality reduction analysis of the LLM's output, we design a penalty function aiming at driving the LLM's outputs away from these abnormal areas as follows:

$$P_{dl} = \sum_{j=1}^{N} (\text{Dist}(x_j, \mathcal{N}) - \text{median}(\text{Dist}(x, \mathcal{N}))), \tag{2}$$

where P_{dl} is the penalty term, x_j represents the PCA-reduced LLM output point in the j-th category of hallucination, $\mathcal{N}$ represents the center of the feature space distribution for correct answers, Dist() is the Euclidean distance function, and N is the number of hallucination categories. Using the median instead of $\mathcal{N}$ as the reference value ensures that points closer to the center have negative penalty values, while points further away have positive penalty values.

Then, we analyze the unusual attention outputs of specific layers or heads when processing different categories of hallucinations, and use the following penalty function to adjust these abnormal attention outputs:

$$P_a = \sum_{k=1}^{N} \sum_{i=1}^{L} \sum_{j=1}^{H} (A_{ijk} - \bar{A}) \cdot \text{sign}(A_{ijk} - \bar{A}), \tag{3}$$

where P_a is the penalty term, L and H represent the number of layers and heads in the LLM, respectively, A_{ijk} is the attention output for the k categories of hallucination in the

i-th layer and j-th head, $\bar{A}$ is the average attention output for correct answers. Sign function is to ensure that the penalty is positive when the attention output value is above the average and negative when it is below the average. Because our datasets all have ground truth, the calculation of the output value of the correct answer such as $\bar{o_i}$ and $\bar{A}$ is enabled. For cases lacking a correct answer, more powerful LLMs like GPT-4 or human can generate them. Additionally, with sufficiently large datasets, a reward model can be trained to score candidate answer probabilities.

After that, based on the correlation coefficient analysis of the attention outputs, we design the following penalty function to close the gap in attention distribution similarity between each type of hallucination and correct outputs:

$$P_{ca} = \sum_{i=1}^{N} \left|1 - \rho(A_i, \bar{A})\right| \cdot \text{sign}(1 - \rho(A_i, \bar{A})), \quad (4)$$

where P_{ca} is the penalty term, N is the number of hallucination categories, A_i is the attention output matrix for the ith category of hallucination, $\bar{A}$ is the average attention output matrix for correct answers, ρ is used to calculate the correlation coefficient between the two matrices, and $||$ is the absolute difference between two outputs. Sign function is to ensure that the penalty is positive when the correlation coefficient difference is greater than 1 and negative when it is less than 1.

Finally, for high-contribution words that play a key role in the generation of hallucinations, we design the following penalty function to reduce the LLM's reliance on these words:

$$P_{wa} = \sum_{i=1}^{N} \sum_{w_i \in W_i} \left(\text{Con}(w_i) - \text{mean}(\text{Con}(W_i))\right), \quad (5)$$

where P_{wa} is the penalty term, W_i is the set of high-contribution words for the ith category of hallucination, and $\text{Con}(w_i)$ represents the contribution of word w_i to generating hallucinatory texts corresponding to a certain category. Take $\text{mean}(\text{Con}(W_i))$ as the baseline is to ensure that the penalty is positive when the contribution value is above the mean and negative when it is below the mean.

By combining all penalty items for each category of hallucination, we obtain the specific total penalty function as follows:

$$P = \lambda_l P_l + \lambda_{dl} P_{dl} + \lambda_a P_a + \lambda_{ca} P_{ca} + \lambda_{wa} P_{wa}, \quad (6)$$

where P is a comprehensive penalty term formed by adding together multiple specific penalty items. λ_l, λ_{dl}, λ_a, λ_{ca} and λ_{wa} are trainable coefficients which are positive values and the sum of them is 1.

Adjusting post-probabilities of candidate answers. After calculating total penalty value P, we adjust the posterior probabilities of each candidate answer. This adjustment process aims to lower the selection probability of those answers containing undesirable hallucination features, while elevating the posterior probabilities of those answers further from hallucination features. The specific adjustment formula is as follows:

$$P_n = P_o \cdot \exp(-\alpha \cdot P), \quad (7)$$

Method	TruthfulQA			RelQA-Cate	
	MC1	MC2	MC3	F1	CGS
LLaMA-7B	23.62	41.21	19.33	32.15	3.39
+Alpaca	26.93	42.97	19.79	36.78	3.76
+ITI	25.90	-	-	37.55	3.98
+CD-13B	24.40	41.00	19.00	34.56	3.59
+DoLa	31.95	52.21	28.17	40.24	4.17
+SH2	27.91	55.63	29.73	39.58	4.06
+DPD	**34.27**	**57.54**	**31.02**	**44.26**	**4.60**
↑	2.32	1.91	1.29	4.02	0.43
↑ (%)	7.26	3.43	4.34	9.99	10.31
LLaMA2-7B	37.62	54.60	28.12	36.52	3.88
+ITI	37.01	54.66	27.82	37.50	4.05
+DoLa	32.97	60.84	29.50	41.28	4.21
+CD-13B	28.15	54.87	29.75	36.89	3.91
+ICD	46.32	69.08	41.25	42.77	4.42
+DPD	**49.63**	**74.28**	**43.99**	**48.55**	**4.73**
↑	3.31	5.20	2.74	5.78	0.31
↑ (%)	7.15	7.53	6.64	13.51	7.01
Baichuan2-7B	34.93	52.14	27.19	38.77	4.28
+ICD	45.75	65.51	39.67	46.83	4.65
+DPD	**50.23**	**68.95**	**44.30**	**50.74**	**4.87**
↑	4.48	3.44	4.63	3.91	0.22
↑ (%)	9.79	5.25	11.67	8.35	4.73
Mistral-7B	39.09	55.80	28.25	40.28	4.53
+ICD	58.53	74.73	50.38	55.31	4.82
+DPD	**60.34**	**78.35**	**52.08**	**58.04**	**4.93**
↑	1.81	3.62	1.70	2.73	0.11
↑ (%)	3.09	4.84	3.37	4.94	2.28

Table 2: Performance of different decoding strategies on several LLMs on TruthfulQA and RelQA-Cate datasets.

where P_n represents the adjusted posterior probability of a candidate answer. P_o is the original posterior probability of candidate answer before penalty adjustment. It involves model probability output that LLMs assign probabilities to possible words or sequences for each candidate answer, as well as is dependent on answer length because longer answers may accumulate more uncertainty, potentially lowering their overall probability. P is the total penalty value of a candidate answer, incorporating all hallucination category-related penalties. α is a global tuning parameter which is a positive value, controlling the impact of the penalty item on the posterior probability.

Experiment

In this section, we conduct extensive experiments to evaluate the positive effect of our proposed DPD in comparison with other decoding strategies.

Experimental setup. Our experiments are conducted on 8xNvidia A100 GPUs, each with 80GB of memory, using PyTorch in Python. We set the maximum sequence length for input and output sequences to maximum 1024 and 128 tokens, respectively. Because the attention patterns of hallucinations differ across datasets. Therefore, our penalties are tailored for each dataset, respectively. Then, we validate the effect in the test set of the same dataset.

Datasets, Metrics, and Baselines. We adopt RelQA-Cate and TruthfulQA (Lin, Hilton, and Evans 2021) datatsets. We utilize F1 score (short for F1) and ChatGPT Score (short

Method	TruthfulQA			RelQA-Cate	
	MC1	MC2	MC3	F1	CGS
LLaMA2-7B	37.62	54.60	28.12	36.52	3.88
+RARR	40.54	61.35	33.13	41.32	4.03
+L	38.93	58.22	30.46	38.97	3.9
+CoVe	44.35	68.13	39.65	43.14	4.25
+CoNLI	42.65	65.36	34.87	42.13	4.3
+RHO	46.66	72.15	42.55	46.87	4.62
+FLEEK	47.76	73.23	42.98	46.75	4.55
+DPD	**49.63**	**74.28**	**43.99**	**48.55**	**4.73**

Table 3: Comparisons with our proposed DPD and other baselines besides decoding strategies for reducing hallucinations.

for CGS) inspired by Chen et al. (2023d), which evaluate the similarity and goodness between the generated answer and the ground-truth answer from RelQA, respectively, on RelQA-Cate. CGS is a 5-scale rating for the generated answer evaluated by ChatGPT with 1 being the worst and 5 being the best for the given question. We use multiple-choice-based metrics including MC1, MC2, MC3 as elaborated in Lin, Hilton, and Evans (2021), which is to evaluate LLMs' performance in TruthfulQA dataset. Baselines are shown in Table 2 and Table 3.

Main Results

As demonstrated in Table 2, the introduction of the DPD has led to improvements across various datasets for each LLM. We conduct a t-test on the results, and all improvements of DPD are statistically significant, with $p < 0.05$. For LLaMA-7B, DPD has brought improvements of up to 7.26% over the previous SOTA strategies on the TruthfulQA dataset and has improved the performance on the RelQA-Cate dataset by 10.31%. With LLaMA2-7B, the DPD also show effective enhancements on the TruthfulQA and RelQA-Cate dataset. Compared with the effect of DPD into Baichuan2-7B, we observe the increases are not as high as those for the LLaMA series, which might be because the LLaMA-7B may contain more hallucinatory phenomena in its original outputs, but Baichuan2-7B may already be optimized for more QA tasks. Similar patterns are also seen in Mistral-7B, suggesting that Mistral-7B may already have a good baseline performance and the addition of DPD provides a subtle accuracy boost. Moreover, we also compare results of non-decoding baselines for reducing hallucinations as shown in Table 3 based on LLaMA2-7B. The results indicate DPD outperforms non-decoding methods across the datasets involved.

Ablation Study

We conduct ablation study on the RelQA-Cate dataset for different LLMs as shown in Fig. 8, Fig. 9, and Table 4, respectively. We first explore candidate answer diversity for DPD. LLMs show the lowest performance without DPD, but as diversity increases from 0 to 0.8, both F1 and CGS scores improve. When diversity exceeds 0.8, indicating DPD, all LLMs achieve optimal performance, which suggests that diversity in candidate answers has positive effect in reducing hallucination in LLMs' outputs. Next, we explore the specific effect of each penalty function in DPD. When P_l is removed, all LLMs experience a performance drop, but this decrease is moderate compared to other penalties. The removal of P_{ca} shows almost no significant performance decrease, as well as the most notable performance degradation occurs when P_{wa} is removed, underscoring the critical role of high-contribution word adjustment in reducing hallucinations. After that, we explore the improvement degree with DPD on different size of LLMs. "↑" and "↑ (%)" indicate the absolute and relative improvements in DPD compared to the methods in the previous row, also proving that our method is effective on larger-sized LLMs.

Method	TruthfulQA			RelQA-Cate	
	MC1	MC2	MC3	F1	CGS
LLaMA-13B	28.55	46.44	26.12	37.12	4.21
+DPD	36.61	58.15	34.23	47.11	4.68
↑	8.06	11.71	8.11	9.99	0.47
↑(%)	**28.23**	**25.22**	**31.05**	**26.91**	**11.16**
LLaMA-33B	37.98	53.91	31.87	45.72	4.43
+DPD	41.24	62.53	36.13	50.38	4.78
↑	3.26	8.62	4.26	4.66	0.35
↑(%)	**8.58**	**15.99**	**13.37**	**0.19**	**7.90**
LLaMA-65B	48.92	67.24	44.30	55.17	4.65
+DPD	52.55	70.14	49.48	59.54	4.83
↑	3.63	2.90	5.18	4.37	0.18
↑(%)	**7.42**	**4.31**	**11.69**	**7.92**	**3.87**
LLaMA2-70B	52.78	74.23	50.25	61.32	4.83
+DPD	57.32	78.39	53.87	64.27	4.86
↑	4.54	4.16	3.62	2.95	0.03
↑(%)	**8.60**	**5.60**	**7.20**	**4.81**	**0.62**
Mistral 8x7B	46.93	63.14	39.18	49.83	4.62
+DPD	61.30	79.15	56.89	63.15	4.97
↑	14.37	16.01	17.71	13.32	0.35
↑(%)	**30.62**	**25.36**	**45.20**	**26.73**	**7.58**

Table 4: The improvement degree of DPD with models of different sizes based on LLaMA on the TruthfulQA and RelQA-Cate datasets.

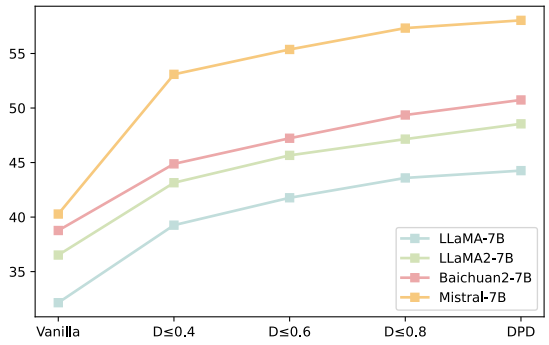

Figure 8: LLMs' performance with different diversity of candidate answers (using F1 score as an example).

Case Study

We conduct a case study to analyze the effectiveness of DPD based on an answer with factual error hallucination as shown

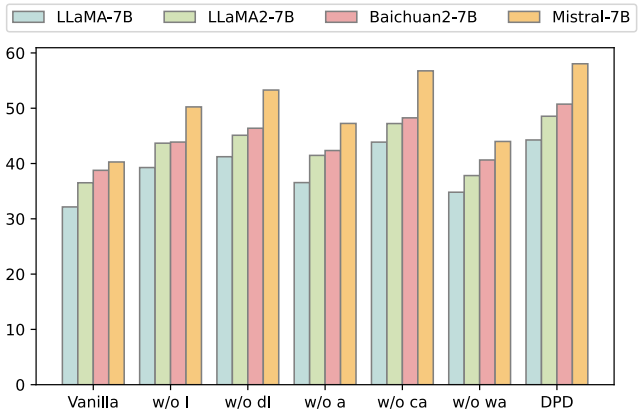

Figure 9: The specific effect of each penalty function in DPD for reducing hallucination (using F1 score as an example).

in Table 5. For the given question "Which is the longest river in Northern Ireland?" with the ground-truth "River Bann", we first generate five candidate answers and obtain their probabilities. Before applying DPD, the predicted answer by the LLM is "The River Lagan." with its probabilities maximum. Next, we utilize the DPD to calculate their penalty values and then adjust the probabilities. We observe that the previous answer lowers its probability and the candidate "The River Bann ... 92 miles" has the maximum probability which is the correct answer. This case demonstrates applying penalties can significantly lower the LLMs' confidence in incorrect answers, effectively reducing hallucinations. Moreover, we also observe different categories of hallucination have different improvements in LLMs' accuracy as shown in Table 6 across samples in Table 1.

Related Work

Reducing hallucination in LLMs. Shi et al. (2023) introduced external knowledge to user queries in the prompt, Peng et al. (2023) enhanced the accuracy of predictions through external knowledge; Zhang et al. (2023) adopted chain-of-thought for guiding LLMs to generate reasoning path, Touvron et al. (2023) added instructions like "If you don't know,..." to guide LLMs not to propagate unverifiable information output; Zhou et al. (2023) manually adjusted data with 1,000 samples annotated by human experts, Cheng et al. (2024) constructed an "I don't know" dataset, training LLMs not to answer questions they don't know; Yu et al. (2023) enhanced the credibility of LLMs through human behavioral adjustment, Sun et al. (2023) proposed factually-augmented RL to enhance the reward model, Chen et al. (2023a) fed LLMs with some incorrect text, making LLMs reflect on the reasons for the errors. However, these methods do not involve with the internal state of LLMs.

Internal state of LLMs. Gurnee and Tegmark (2023) discovered that LLMs can learn linear representations of space and time across multiple spacetime scales; Chen et al. (2023e) proposed "Attention Buckets" based on RoPE for each attention module; Chen et al. (2023c) analyzed the attention output of the Hadamard adapter and full fine-tuning are similar in performance; Ziheng et al. (2023) constructed a set of globally shared adjustable tokens to modify the attention of each layer for LLMs; Xu et al. (2023) analyzed the relative token contributions to model's generation. CH-Wang et al. (2023) developed probes trained on transformer model representations in in-context generation tasks. Inspired by the above methods, we also analyze LLMs' internal state to guide the method design.

Candidate answers	Probability	Penalty	AdjustedP
The River Bann in County Antrim is the longest river in Northern Ireland with a course of 92 miles	0.62	0.17	0.52
The River Liffey, at 110 miles, is Ireland's longest river	0.58	0.44	0.37
The Longest River in Ireland is the River Shannon	0.63	0.67	0.32
The River Lagan	0.69	0.75	0.32
The River Foyle is the longest river in Northern Ireland	0.55	0.73	0.28

Table 5: Candidate answers with the original probabilities, the penalty values obtained with DPD, and the adjusted probabilities for a given question.

Category	Probability	Penalty	AdjustedP
Factual error	0.65	0.82	0.29
Logical error	0.67	0.73	0.32
Conceptual confusion	0.73	0.97	0.28
Vagueness	0.66	0.59	0.37
Lack of commonsense	0.58	0.43	0.38
Over-generalization	0.89	0.65	0.46
Emotional bias	0.65	0.9	0.26
Lack of uncertainty	0.72	0.85	0.31

Table 6: Different categories of hallucination examples corresponding to Table 1 with the original probabilities, DPD generated penalty values and adjusted probabilities.

Conclusions and Future Work

Large language models (LLMs) currently demonstrate excellent capabilities in a variety of downstream tasks, but the hallucinations of the output answers pose serious challenges. In this paper, we propose an attribution framework to trace the origins of hallucinations based on the internal signals of LLMs. Further, we propose a novel Differential Penalty Decoding (DPD) strategy to assign penalty values to each answer with hallucination and adjust the post-probabilities of these answers, making the hallucination output less likely to be selected. Our experiments demonstrates that DPD performs well on various datasets and LLMs, making significant contributions in mitigating hallucinations of LLMs.

References

Cao, Y.; Kang, Y.; and Sun, L. 2023. Instruction mining: High-quality instruction data selection for large language models. *arXiv preprint arXiv:2307.06290*.

CH-Wang, S.; Van Durme, B.; Eisner, J.; and Kedzie, C. 2023. Do Androids Know They're Only Dreaming of Electric Sheep? *arXiv preprint arXiv:2312.17249*.

Chen, K.; Wang, C.; Yang, K.; Han, J.; Hong, L.; Mi, F.; Xu, H.; Liu, Z.; Huang, W.; Li, Z.; et al. 2023a. Gaining wisdom from setbacks: Aligning large language models via mistake analysis. *arXiv preprint arXiv:2310.10477*.

Chen, L.; Li, S.; Yan, J.; Wang, H.; Gunaratna, K.; Yadav, V.; Tang, Z.; Srinivasan, V.; Zhou, T.; Huang, H.; et al. 2023b. Alpagasus: Training a better alpaca with fewer data. *arXiv preprint arXiv:2307.08701*.

Chen, Y.; Fu, Q.; Fan, G.; Du, L.; Lou, J.-G.; Han, S.; Zhang, D.; Li, Z.; and Xiao, Y. 2023c. Hadamard Adapter: An Extreme Parameter-Efficient Adapter Tuning Method for Pre-trained Language Models. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 276–285.

Chen, Y.; Fu, Q.; Yuan, Y.; Wen, Z.; Fan, G.; Liu, D.; Zhang, D.; Li, Z.; and Xiao, Y. 2023d. Hallucination detection: Robustly discerning reliable answers in large language models. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 245–255.

Chen, Y.; Li, Y.; Yan, S.; Liu, S.; Liang, J.; and Xiao, Y. 2024a. Do Large Language Models have Problem-Solving Capability under Incomplete Information Scenarios? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Chen, Y.; Lv, A.; Lin, T.-E.; Chen, C.; Wu, Y.; Huang, F.; Li, Y.; and Yan, R. 2023e. Fortify the Shortest Stave in Attention: Enhancing Context Awareness of Large Language Models for Effective Tool Use. *arXiv preprint arXiv:2312.04455*.

Chen, Y.; Wen, Z.; Fan, G.; Chen, Z.; Wu, W.; Liu, D.; Li, Z.; Liu, B.; and Xiao, Y. 2023f. Mapo: Boosting large language model performance with model-adaptive prompt optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 3279–3304.

Chen, Y.; Xiao, Y.; Li, Z.; and Liu, B. 2023g. XMQAs: Constructing Complex-Modified Question-Answering Dataset for Robust Question Understanding. *IEEE Transactions on Knowledge and Data Engineering*.

Chen, Y.; Xiao, Y.; and Liu, B. 2022. Grow-and-Clip: Informative-yet-Concise Evidence Distillation for Answer Explanation. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, 741–754. IEEE.

Chen, Y.; Yan, S.; Guo, Q.; Jia, J.; Li, Z.; and Xiao, Y. 2024b. HOTVCOM: Generating Buzzworthy Comments for Videos. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Chen, Y.; Yan, S.; Liu, P.; and Xiao, Y. 2024c. Dr.Academy: A Benchmark for Evaluating Questioning Capability in Education for Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Chen, Y.; Yan, S.; Liu, S.; Li, Y.; and Xiao, Y. 2024d. EmotionQueen: A Benchmark for Evaluating Empathy of Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Chen, Y.; Yan, S.; Zhu, Z.; Li, Z.; and Xiao, Y. 2024e. XMeCap: Meme Caption Generation with Sub-Image Adaptability. In *Proceedings of the 32nd ACM Multimedia*.

Chen, Y.; Yuan, Y.; Liu, P.; Liu, D.; Guan, Q.; Guo, M.; Peng, H.; Liu, B.; Li, Z.; and Xiao, Y. 2024f. Talk Funny! A Large-Scale Humor Response Dataset with Chain-of-Humor Interpretation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 17826–17834.

Chen, Y.; Zhao, J.; Wen, Z.; Li, Z.; and Xiao, Y. 2024g. TemporalMed: Advancing Medical Dialogues with Time-Aware Responses in Large Language Models. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, 116–124.

Cheng, Q.; Sun, T.; Liu, X.; Zhang, W.; Yin, Z.; Li, S.; Li, L.; Chen, K.; and Qiu, X. 2024. Can AI Assistants Know What They Don't Know? *arXiv preprint arXiv:2401.13275*.

Gurnee, W.; and Tegmark, M. 2023. Language models represent space and time. *arXiv preprint arXiv:2310.02207*.

Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12): 1–38.

Li, J.; Chen, J.; Ren, R.; Cheng, X.; Zhao, W. X.; Nie, J.-Y.; and Wen, J.-R. 2024. The Dawn After the Dark: An Empirical Study on Factuality Hallucination in Large Language Models. *arXiv preprint arXiv:2401.03205*.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Lin, S.; Hilton, J.; and Evans, O. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Peng, B.; Galley, M.; He, P.; Cheng, H.; Xie, Y.; Hu, Y.; Huang, Q.; Liden, L.; Yu, Z.; Chen, W.; et al. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.

Shi, W.; Min, S.; Yasunaga, M.; Seo, M.; James, R.; Lewis, M.; Zettlemoyer, L.; and Yih, W.-t. 2023. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Sun, Z.; Shen, S.; Cao, S.; Liu, H.; Li, C.; Shen, Y.; Gan, C.; Gui, L.-Y.; Wang, Y.-X.; Yang, Y.; et al. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

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

Xu, W.; Agrawal, S.; Briakou, E.; Martindale, M. J.; and Carpuat, M. 2023. Understanding and detecting hallucinations in neural machine translation via model introspection. *Transactions of the Association for Computational Linguistics*, 11: 546–564.

Yu, T.; Yao, Y.; Zhang, H.; He, T.; Han, Y.; Cui, G.; Hu, J.; Liu, Z.; Zheng, H.-T.; Sun, M.; et al. 2023. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *arXiv preprint arXiv:2312.00849*.

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

Zhong, M.; Yin, D.; Yu, T.; Zaidi, A.; Mutuma, M.; Jha, R.; Awadallah, A. H.; Celikyilmaz, A.; Liu, Y.; Qiu, X.; et al. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. *arXiv preprint arXiv:2104.05938*.

Zhou, C.; Liu, P.; Xu, P.; Iyer, S.; Sun, J.; Mao, Y.; Ma, X.; Efrat, A.; Yu, P.; Yu, L.; et al. 2023. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*.

Ziheng, Z.; Wu, Y.; Zhu, S.-C.; and Terzopoulos, D. 2023. Aligner: One Global Token is Worth Millions of Parameters When Aligning Large Language Models. *arXiv preprint arXiv:2312.05503*.