Focus on Your Question! Interpreting and Mitigating Toxic CoT Problems in Commonsense Reasoning

Jiachun Li[1,2], Pengfei Cao[1,2], Chenhao Wang[1,2], Zhuoran Jin[1,2], Yubo Chen[1,2,*]
Daojian Zeng[3], Kang Liu[1,2], Jun Zhao[1,2,*]

[1]School of Artificial Intelligence, University of Chinese Academy of Sciences
[2]The Laboratory of Cognition and Decision Intelligence for Complex Systems,
Institute of Automation, Chinese Academy of Sciences
[3]Hunan Normal University

{jiachun.li, pengfei.cao, chenhao.wang, zhuoran.jin, yubo.chen, kliu, jzhao} @nlpr.ia.ac.cn
zengdj916@163.com

Abstract

Large language models exhibit high-level commonsense reasoning abilities, especially with enhancement methods like Chain-of-Thought (CoT). However, we find these CoT-like methods lead to a considerable number of originally correct answers turning wrong, which we define as the Toxic CoT problem. To interpret and mitigate this problem, we first utilize attribution tracing and causal tracing methods to probe the internal working mechanism of the LLM during CoT reasoning. Through comparisons, we prove that the model exhibits information loss from the question in the shallow attention layers when generating rationales or answers. Based on the probing results, we design a novel method called $\mathbb{RIDERS}$ (**R**esidual decod**I**ng and s**ER**ial-position **S**wap), which compensates for the information deficit in the model from both decoding and serial-position perspectives. Through extensive experiments on multiple commonsense reasoning benchmarks, we validate that this method not only significantly eliminates Toxic CoT problems (decreased by **23.6%**), but also effectively improves the model's overall commonsense reasoning performance (increased by **5.5%**).

1 Introduction

With the increase in scale, large language models (LLMs) have demonstrated outstanding performance in different tasks (Li et al., 2023; Wen et al., 2023; Sun et al., 2024; Jin et al., 2024a), among them, commonsense reasoning has received significant attention due to its importance for general intelligence (Wang et al., 2022, 2024; Liu et al., 2024). In this task, researchers have proposed a series of chain-of-thought (CoT) like techniques to elicit models' potential abilities (e.g. Self-Consistency (Wang et al., 2023c), Least-to-Most (Zhou et al., 2023), Reflexion (Shinn et al., 2023)).
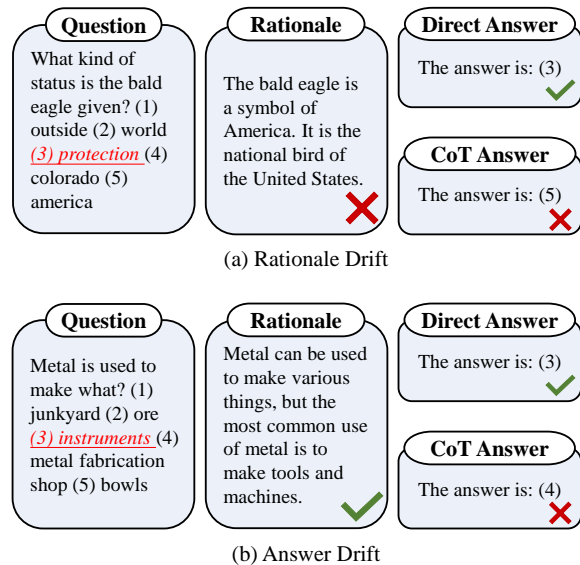

Figure 1: Two examples for the Toxic CoT problem.

Through them, LLMs can generate reasonable rationales and improve their reasoning performance.

While these works have made great progress, we notice an overlooked problem in them, which we define as **Toxic CoT** — Sometimes LLMs can directly provide correct answers to questions, but after applying CoT-like methods, it brings extra reasoning paths to models, causing their answers to be wrong.[1] Figure 1 illustrates two main error types of this problem — Rationale Drift and Answer Drift. Specifically, for the Rationale Drift case, given the question "*What kind of status is the bald eagle given?*", the model can directly give the correct answer "*protection*". However, in the rationale, the model explains "*what is the bald eagle*" as "*a symbol of America*", which has a semantic drift from the question. Thus, the model chooses the wrong option "*america*" based on the drifting rationale. For the Answer Drift case, given the question

---

*Corresponding authors.

[1]Notably, our definition here differs from the toxicity in generated content (Shaikh et al., 2023), emphasizing the harm that CoT can bring to the model.

"*Metal is used to make what?*", the model can directly answer "*instruments*". It can also generate a correct rationale "*metal is to make tools and machines*", but when answering based on the rationale, the model drifts from it and selects the incorrect option "*(4)*". We further conduct a statistical analysis over extensive commonsense reasoning datasets and find that, among all CoT errors, this problem accounts for **37%** for the white-box model and **33%** for the black-box model on average, indicating this problem has become a crucial bottleneck in CoT reasoning.[2]

So what is the mechanism behind this issue? In this paper, we attempt to answer this question by probing the inner workings of the LLM's CoT reasoning. Specifically, we first make initial observations on examples of Rationale Drift and Answer Drift issues, which suggest that the model likely misses some important information from the question when generating corresponding rationales or answers. To further verify these findings, we use attribution tracing and causal tracing methods to probe the LLM in two stages (rationale generation stage and answer generation stage). By employing these methods under various experimental settings, **we find that there is a significant loss of information flow from the question in the shallow attention layers when generating drifting rationales and answers.** Therefore, we interpret the Toxic CoT problem as the model lacking information from the question in the two stages.

To validate our interpretation and mitigate this problem, we design an approach called ℝ𝕀𝔻𝔼ℝ𝕊 (**R**esidual decod**I**ng and s**ER**ial-position **S**wap) based on the interpretation. Concretely, for the Rationale Drift issue, we devise a decoding algorithm, promoting the model to generate tokens that pay more attention to question contexts. For the Answer Drift issue, we swap the positions of the output sequence, reducing the information loss from the question to the final prediction. We evaluate our method on five commonsense reasoning benchmarks and conduct extensive experiments. The results not only prove our interpretation, but also indicate that our method is effective in addressing the Toxic CoT problem and improving the model's overall commonsense reasoning abilities.

We summarize the contribution of this paper as follows:

(1) We identify a crucial bottleneck affecting LLM's reasoning performance called the Toxic CoT problem, probe this issue through attribution tracing and causal tracing methods, and interpret the mechanism behind it as the model missing information from questions in shallow attention layers. The results contribute to a more in-depth understanding of the LLM's reasoning mechanisms.

(2) To mitigate the Toxic CoT problem, we introduce ℝ𝕀𝔻𝔼ℝ𝕊, which effectively compensates for the internal information loss during CoT reasoning from decoding and serial-position perspectives.

(3) We conduct extensive experiments on various benchmarks. The results not only verify the rationality of our interpretation, but also demonstrate the effectiveness of our method in addressing the Toxic CoT problem (the proportion of the problem decreased by **23.6%**) and enhancing commonsense reasoning performance (overall accuracy increased by **5.5%**). Our code is available at: https://github.com/BugMakerzzz/toxic_cot.

2 Problem Statement

2.1 Toxic Chain of Thought Reasoning

We start our work by formally defining the Toxic CoT problem as follows: [3]

Definition 2.1 (**Toxic CoT**). *Given a question q and the correct answer o^*, if the model's output $\mathcal{M}$ meets the following conditions, it is considered a case of Toxic CoT:*

$$o^* = \mathcal{M}(q, P_d) \wedge o^* \neq \mathcal{M}(q, P_c)$$

where $o^* = \mathcal{M}(q, P_d)$ indicates the model's direct answering for q is correct, $o^* \neq \mathcal{M}(q, P_c)$ indicates the model's cot-like answering for q is wrong, P_d, P_c are the corresponding prompts.

2.2 Two-stage Drift Issues

To investigate the reasons for the problem, we classify these Toxic CoT cases and identify a main error causing this problem (On average, it accounts for 67% on two datasets, see more details in Appendix B). Furthermore, if we divide the CoT process into two stages: **rationale generation** and **answer generation**, there exist two types of issues in this error:

Definition 2.2 (**Rationale Drift**). *If the reasoning chain is factually correct but logically inconsistent*

---

with the question, this case is called "Rationale Drift" (see Figure 1a).

Definition 2.3 (**Answer Drift**). *If the reasoning chain is both factually correct and logically consistent with the question, but the final answer is inconsistent with the rationale, this case is called "Answer Drift" (see Figure 1b).*

2.3 Hypothesis Formulation

To provide a direction for subsequent probing experiments, here we attempt to propose hypotheses for the mechanism of issues by analyzing some examples. For the Rationale Drift issue, the model tends to focus on part of the essential reasoning conditions in the question context. As an example, in Figure 1a, the CoT only focuses on the "*bald eagle*" in the question but misses another key information "*status*". As for the Answer Drift issue, the model seems to be disrupted by CoT, losing attention to the question and resulting in an off-topic prediction. For instance, in Figure 1b, though the CoT gives correct information "*to make tools and machines*", the model can only predict the wrong answer "*(4) metal fabrication shop*". This is likely because the model loses the question's target "*to make what*" and directly copies the entity "*metal*", which frequently appears in CoT, as the answer. Therefore, we summarize our hypotheses as follows:

Hypothesis 1. *The Rationale Drift issue arises from the model lacking information from the question context in the rationale generation stage.*

Hypothesis 2. *The Answer Drift issue arises from the model lacking information from the question in the answer generation stage.*

To validate the above hypotheses, in the following two sections, we conduct probing experiments, exploring the LLM's internal working mechanisms during the two stages of CoT reasoning.

3 Tracing Information Flow in Rationale

In this section, we aim to verify the Hypothesis 1 by tracing the information flow in the rationale generation stage. To this end, we start by describing our attribution tracing method (§3.1). Through this method, we conduct comparative experiments between the correct reasoning and the drifting one, figuring out the mechanism behind the issue (§3.2). At last, we use the attention score to validate our findings from another perspective (§3.3).
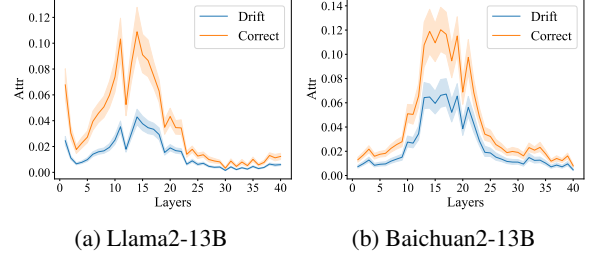

(a) Llama2-13B (b) Baichuan2-13B

Figure 2: Attribution tracing results on Winogrande.

3.1 Tracing Method

To investigate the roles of different model components in the rationale generation stage, we use attribution scores (Hao et al., 2021; Dai et al., 2022; Wang et al., 2023b) to compute the contribution of a neuron ω:

$$Attr(\omega) = \omega \odot \int_{\alpha=0}^{1} \frac{\partial F(\alpha\omega)}{\partial \omega} d\alpha \approx \frac{\omega}{m} \odot \sum_{k=1}^{m} \frac{\partial F(\frac{k}{m}\omega)}{\partial \omega} \tag{1}$$

where $F(\cdot)$ represents the model's output. We compute the attribution score via Riemman approximation of the integration and m is the number of approximation steps. For neurons $A^{(l)}$ in the i-th attention layer, we sum the absolute values of scores on all attention heads to get the final attribution score. Since the attention module involves interactions between different tokens, we can compute the information flow between the question context q and the CoT c on it:

$$Q_{qc}^{(l)} = \frac{1}{|N|} \sum_{(i,j) \in C_{qc}} Attr(A_{i,j}^{(l)})$$
$$C_{qc} = \{(i,j) | q_s \leq i \leq q_e,\ c_s \leq j \leq c_e\} \tag{2}$$

Here, $Attr(A_{i,j}^{(l)})$ represents the intensity of information flow from the i-th token to the j-th token in the l-th attention layer and $|N|$ denotes the number of CoT steps. More implementation details of this method are reported in Appendix C.

3.2 Attribution Tracing Experiment

Experimental Settings To validate the deficiency in Hypothesis 1, we need to figure out the context information flow difference between generating a drifting rationale and a correct one. Thus, we first use golden labels as hints to generate correct CoTs in the drifting cases. Then, we compute the average information flow under these two cases and compare their results. We choose Llama2-13B (Touvron et al., 2023) and Baichuan2-13B (Yang et al., 2023) as our probing models, since they are moderated-sized white-box models with decent
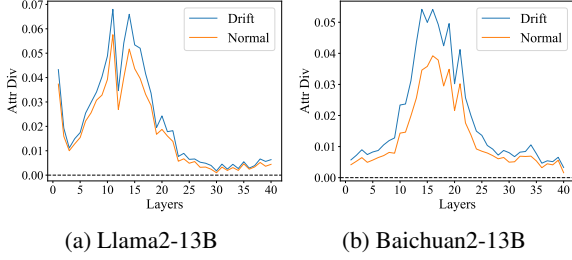
Figure 3: Information flow divergence comparison on Winogrande.

Figure 4: Attention tracing results across different attention heads on Llama2-13B.

CoT performance. For datasets, we select Winogrande (Sakaguchi et al., 2020) and CSQA (Talmor et al., 2019).[4] The detailed implementation of this experiment is shown in Appendix C.

Result and Analysis   Figure 2 illustrates our experimental results on Winogrande (The results on CSQA are shown in Appendix C). We can find that: **(Claim 1) When the Rationale Drift issue occurs, CoT receives less information from the question context compared to the correct case.** On both datasets and different models, there is a significantly lower information flow between the question context and CoT when the LLM generates a drifting rationale (the blue line) compared to the correct one (the orange line). This aligns with Hypothesis 1. **(Claim 2) The shallow attention layers are crucial for LLMs to extract contextual information.** In all cases, both the information flow and the gap peak at around the 15th attention layer, indicating these layers are significant sites for the rationale generation.

Supplementary Experiment   In the main experiment, we use golden labels to generate the correct CoT. To eliminate the influence of this additional factor on our results, we design a supplementary experiment. Concretely, we first use the label to generate CoTs from correct reasoning cases, which serves as a control group. Then, we compute the context information divergence between the newly generated CoT c_n and the original one c_o, i.e.:

$$Attr(c_n|c_o) = Q_{qc_n}^{(l)} - Q_{qc_o}^{(l)} \qquad (3)$$

where q is the question context. We compare this divergence between the control and drifting group, whose results are reported in Figure 3 and Figure 10. As we can see, the gap between the correct CoT and the drifting one (the blue line) is larger

than the control group (the orange line) and the max divergence occurs in shallow layers (around the 15th layer). **This indicates that a correct CoT indeed gets more information flow from the context in shallow attention layers, validating the effectiveness of Claim 1 and 2.**

3.3   Attention Tracing Experiment

Experimental Settings   We also design an experiment based on attention scores to validate Hypothesis 1 from another perspective. For a pair of rationales $<c, c^*>$ targeting the same question context q (c denotes the correct CoT and c^* is the drifting one), we compute their attention divergence:

$$Attn(c|c^*) = \sum_{(i,j) \in C_{qc}} \frac{A_{i,j}^{(l,h)}}{|c|} - \sum_{(i,j) \in C_{qc^*}} \frac{A_{i,j}^{(l,h)}}{|c^*|} \qquad (4)$$

Here, we replace the $Attr(A_{i,j}^{(l)})$ in Equation 2 with the weights on the h-th attention matrix head $A_{i,j}^{(l,h)}$ and repeat the calculation in Equation 3.

Result and Analysis   The results on two datasets are shown in Figure 4 and Figure 11. We can get the following observations: (1) The attention divergence is greater than 0 in most heads, which indicates **a lack of information from the question context in Rationale Drift cases (consistent with Claim 1).** (2) **The largest attention divergence appears around layer 15, which is consistent with the sites we find in Claim 2.** This once again illustrates that attention heads of these layers are crucial for the LLM to obtain contextual information when generating CoT.

4   Tracing Information Flow in Answer

In this section, our goal is to verify the information loss based on Hypothesis 2. To achieve this, we first introduce the main tracing method in this section, which is called the causal tracing method

---

[4]Unless otherwise specified, we use the same models and datasets in the following probing experiments.

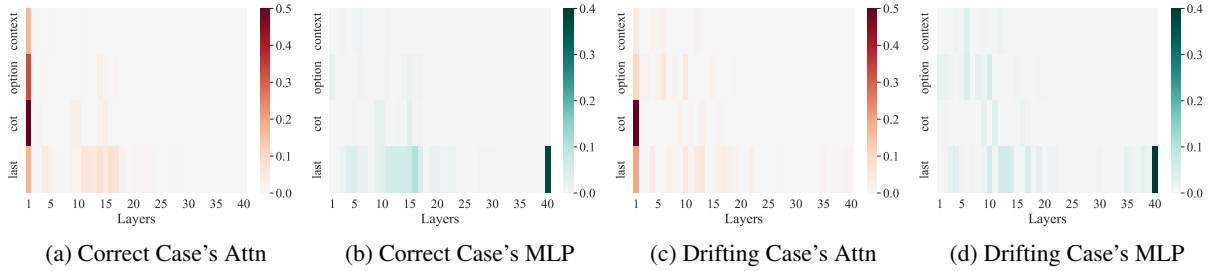|			
(a) Correct Case's Attn	(b) Correct Case's MLP	(c) Drifting Case's Attn	(d) Drifting Case's MLP

Figure 5: Intervention tracing results on Winogrande in correct and drifting answering cases.

(§4.1). Next, by employing it, we trace the information flow in the answer generation stage and identify the mechanism behind the Answer Drift issue through comparative experiments (§4.2). At last, we apply the attribution tracing method to verify our hypothesis from another perspective (§4.3).

4.1 Tracing Method

Since the task we study is in the form of multiple-choice questions, we set our focus on the feedforward pass that predicts the label. Inspired by the previous works (Meng et al., 2022; Stolfo et al., 2023; Geva et al., 2023), we take the causal tracing method to quantify the contribution of different intermediate variables during this pass. Specifically, for hidden states $h_i^{(l)}$ in a clean run that predicts the answer, we have:

$$h_i^{(l)} = h_i^{(l-1)} + a_i^{(l)} + m_i^{(l)}$$
$$a_i^{(l)} = attn^{(l)}(h_1^{(l-1)}, ..., h_i^{(l-1)}) \qquad (5)$$
$$m_i^{(l)} = mlp^{(l)}(a_i^{(l)} + h_i^{(l-1)})$$

where i, l is the i-th token in the l-th layer, $a_i^{(l)}, m_i^{(l)}$ represents the activations of attention and MLP modules in Transformer (Vaswani et al., 2017). Supposing that a certain input part is represented as $z = [v_i^{(l)}, ..., v_j^{(l)}]$ after passing through a model component, we set $v_k^{(l)} = v_k^{(l)} + \epsilon$ for $k \in [i, j]$ to intervene this hidden vector, where ϵ is Gaussian noise.[5] Thus, we can compute the direct effect (DE) of this component:

$$DE(z) = \frac{P(o) - P_z^*(o)}{P(o)} \qquad (6)$$

where $P(o)$ is the probability of the model's final prediction, $P_z^*(o)$ is the probability after the intervention. Therefore, through this metric, we can quantify the contributions of different components in changing the final prediction, thereby tracing the information flow in this stage.

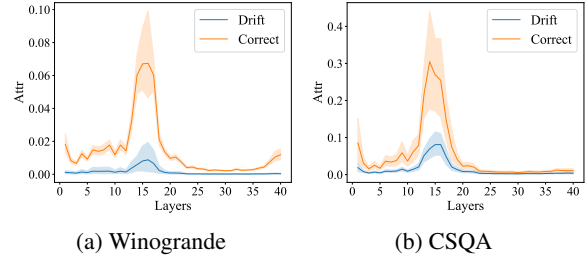|	
(a) Winogrande	(b) CSQA

Figure 6: Attribution tracing results on Llama2-13B during the answer generation stage.

4.2 Intervention Tracing Experiment

Experimental Settings We sample correct and drifting answering cases from datasets, average over them and compute the average direct effect (ADE). Here we compute the impact of four components on the final prediction: context (question contexts), option (question options), CoT, and last (the last token before the label prediction).

Results and Analysis We report the result of Llama2-13B on Winogrande in Figure 5 and others in Appendix D, from which we can get two conclusions: **(Claim 3) For attention modules, drifting cases loss information from the question.** When the answer is correct, we can observe a high effect on the context and option in the first layer (see Figure 5a). But for the drifting case, the LLM extracts limited information at these positions (see Figure 5c). This aligns with Hypothesis 2. **(Claim 4) For MLP modules, the information is not lost.** We observe the same high-effect sites in the last layer and shallow layers of the last token, they do not show regular differences (see Figure 5b and 5d).

4.3 Attribution Tracing Experiment

Experimental Settings For further validation of our hypothesis, we also use the attribution score in §3 to trace the information flow in this stage. Referring to Equation 2, we compute the score between the question context and the last token
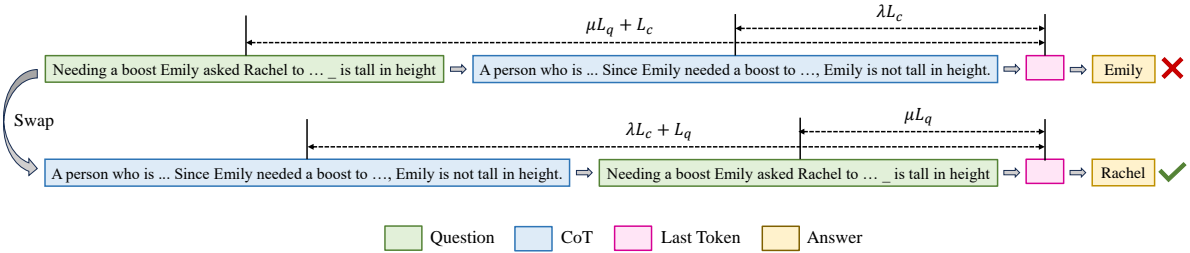
---

[5]We select ϵ to be 3 times larger than the empirical standard deviation of hidden embeddings in each dataset.

Figure 7: An example of our serial-position swap method.

Results and Analysis The results of this experiment are reported in Figure 6 and 15. We can observe that, **when the Answer Drift issue occurs, the information flow from the question significantly decreases**. This verifies the information loss we mention in Hypothesis 2 and Claim 3.

(since it's used for generating the answer). We set the $F(\cdot)$ in Equation 1 as the loss for predicting the final answer, comparing the scores for correct and drifting cases after averaging across samples.

5 Mitigating Toxic CoT Problem

In this section, we propose a novel method called $\mathbb{RIDERS}$ (**R**esidual decod**I**ng and s**ER**ial-position **S**wap) to address the Toxic CoT problem. We first introduce the two components in it, which are designed based on Hypothesis 1 and 2, respectively (§5.1). Then, we conduct experiments on commonsense reasoning benchmarks, demonstrating the effectiveness of our method (§5.2 and §5.3). At last, we conduct extra experiments to further emphasize the contribution of our approach (§5.4).

5.1 Mitigation Method

Residual Decoding We design a new decoding methodology to address the Rationale Drift issue, in which we construct a virtual residual structure during the CoT generation, "connecting" the question context with each CoT token. Our decoding algorithm is demonstrated in Algorithm 1. In each iteration of generating a new token, we first select the top n tokens with the highest probabilities and record their logits scores (line 3). Then we calculate the attention score between the context and current token like Equation 4, normalize it, and add it as an additional reward to promote more information flow (lines 6,7). Finally, we select the token with the highest score to update the input and repeat the process until the termination condition is met. We use the attention matrix in layer 15 to

---

Algorithm 1 Residual Decoding Algorithm

Require: model $\mathcal{M}$, input x, question context q, candidate_num n, weight ω.
1: **for** iteration $i \in 0, 1, ...$ **do**
2:    $logits = \mathcal{M}(x)$
3:    $tokens, scores = \text{top}(logits, n)$
4:    **for** $j \in 1, ...n$ **do**
5:       $t = tokens[i]$
6:       $attn\_score = \text{Attn}(q, t) \,/\, \text{Attn}(x, t)$
7:       $scores[j] = scores[j] + \omega * attn\_score$
8:    **end for**
9:    $idx = \text{argmax}(scores)$
10:    $t = tokens[idx]$
11:    $x = x + t$
12:    **if** $\text{stop}(t)$ **then**
13:       **break**
14:    **end if**
15: **end for**
16: **return** x

---

compute the attention score, since it is crucial for the exchange of contextual information according to Claim 2. More implementation details of this method are provided in Appendix E.

Serial-Position Swap In this method, we attempt to compensate for the information lack in the Answer Drift issue. According to previous research on the serial-position effect in context, models tend to utilize information better at the beginning and end of the input (Qin et al., 2023; Liu et al., 2023b). In our work, the beginning of the input are prompts, while the current question and the generated CoT are both located at the end. Therefore, when they are closer to the last token, their information is more easily utilized in the final prediction. As in Figure 7, we denote the lengths of the question and CoT as L_q and L_c, and assume that the key information is located at positions μL_q and λL_c (similar to the center of mass in physics). We can infer that, after the swapping operation in Figure 7, the distance from the question to the end is reduced ($\mu L_q + L_c \rightarrow \mu L_q$). Besides, if we consider the total distance from the question and CoT to the end, we can perform the following calculation:

Method	Winogrande		CSQA		HellaSwag		SIOA		PIQA		Avg	
	ACC↑	TR↓	ACC↑	TR↓	ACC↑	TR↓	ACC↑	TR↓	ACC↑	TR↓	ACC↑	TR↓
Few-shot Answer	57.1	-	66.8	-	45.3	-	68.2	-	61.6	-	59.8	-
Chain-of-Thought	56.7	43.1	69.9	28.8	45.4	34.8	67.3	37.3	62.7	41.3	60.4	37.1
Self-Consistency	56.4	40.9	72.4	24.9	46.2	33.7	64.7	41.0	54.6	45.8	58.9	37.3
Self-Refine	48.7	50.3	55.5	47.0	42.0	37.2	65.1	38.8	50.8	48.3	52.4	44.3
Least-to-Most	58.9	38.8	69.0	27.5	31.3	39.4	66.3	36.6	65.7	32.2	58.2	34.9
Contrasive CoT	59.4	38.0	71.1	23.2	45.9	35.6	67.0	36.0	68.1	44.9	62.3	35.5
Ours RD Only	58.6	29.5	72.2	13.5	49.4	24.0	69.4	20.8	66.3	26.2	63.0	22.8
SPS Only	59.1	22.8	72.6	18.9	49.3	18.7	69.7	28.8	68.5	27.6	63.7	23.4
RIDERS	**60.7**	**16.6**	**73.2**	**6.7**	**50.6**	**12.8**	**71.9**	**15.3**	**69.9**	**15.9**	**65.3**	**13.5**

Table 1: Performance comparison across five commonsense reasoning datasets on Llama2-13B.

$$d_1 = \mu L_q + L_c + \lambda L_c$$
$$d_2 = \lambda L_c + L_q + \mu L_q \qquad (7)$$
$$\Delta d = d_2 - d_1 = L_q - L_c$$

where d_1 is the total distance in normal serial positions and d_2 is the distance after swapping the two components. In most scenarios, we have $L_q < L_c$, thus, we can infer that $\Delta d < 0$. That means, if we replace the original order of "*[Question] + [CoT]*" with the order of "*[CoT] + [Question]*", we can not only increase the intensify of information flow from the question to the final prediction, but also reduce the total information loss due to the reduction in total distance. Although this method is straightforward in implementation, it proves to be effective in both theory and experiments.

5.2 Mitigation Experimental Settings

Datasets Following previous works, we use five representative commonsense reasoning benchmarks: **WinoGrande** (Sakaguchi et al., 2020), **CSQA** (Talmor et al., 2019), **HellaSwag** (Zellers et al., 2019), **SIQA** (Sap et al., 2019) and **PIQA** (Bisk et al., 2020). The specific information of each dataset is reported in Appendix F.

Metrics In addition to the commonly used **Accuracy (ACC)** metric, we also introduce a new metric — **Toxic Rate (TR)**, to quantify the severity of Toxic CoT problems:

$$TR(f) = |C_d \cap W_f| / |W_f| \qquad (8)$$

where C_d denotes questions that models give correct answers directly and W_f denotes questions that models give wrong answers after applying method f. Thus, we can infer that the lower the TR, the fewer Toxic CoT problems the method introduces.

Baselines As our research focuses on enhancing CoT methods in commonsense reasoning, we

Method	Winogrande		CSQA	
	Type1	Type2	Type1	Type2
CoT	0.0	0.0	0.0	0.0
RD Only	41.9	47.8	56.7	46.2
SPS Only	58.1	78.3	46.7	**92.3**
RIDERS	**74.2**	**82.6**	**73.3**	84.6

Table 2: Accuracy on the two types of drifting issues.

select some of the latest CoT-like methods applicable to this task for comparison: **Few-shot Answer**, **Chain-of-Thought** (Wei et al., 2022), **Self-Consistency** (Wang et al., 2023c), **Self-Refine** (Madaan et al., 2023), **Least-to-Most** (Zhou et al., 2023) and **Contrasive CoT** (Chia et al., 2023). For all methods, we employ a 5-shot prompt and use 4 NVIDIA GeForce RTX 3090 GPUs for inference. More implementation details can be found in the Appendix F.

5.3 Mitigation Results

The main result of our experiments on Llama2-13B is shown in Table 1. (The results on more models are presented in Appendix F.) We can get the following conclusions: **(1) Our method effectively mitigates Toxic CoT problems.** Compared to CoT prompting, our method reduces the Toxic Rate by an average of **23.6%** across five datasets. Besides, compared to other advanced CoT-like methods, our method causes the fewest Toxic CoT problems (decreased by an average of **22.0%** over SOTA methods). **(2) Our method can also improve the model's overall performance on commonsense reasoning.** Our work improves the accuracy on all benchmarks (improved by **5.5%** compared to CoT and **3.0%** compared to SOTA methods on average). This proves that the Toxic CoT problem poses a bottleneck in LLM's commonsense reasoning, highlighting the value of our work.
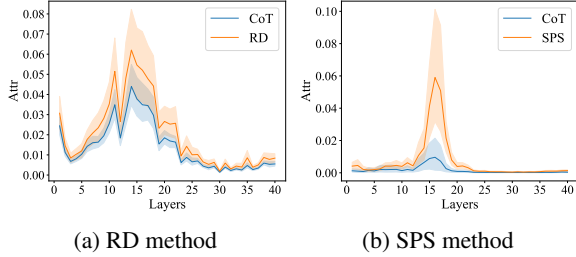
(a) RD method   (b) SPS method

Figure 8: Information flow comparison on Winogrande after applying our two methods.

Method	ProofWriter		FOLIO	
	ACC↑	TR↓	ACC↑	TR↓
CoT	56.8	19.9	45.1	11.6
RIDERS	**78.2**	**6.4**	**48.0**	**7.5**

Table 3: Performance comparison across two logical reasoning tasks on Llama2-13B.

5.4 Discussion and Analysis

Performance on Two Drifting Issues To demonstrate the effectiveness of our method in addressing the Rationale Drift issue (Type1) and Answer Drift issue (Type2), we conduct experiments on these samples and report the results in Table 2. Both of our methods can mitigate the corresponding issues (RD solves **49.3%** Rationale Drift issue on average, while SPS resolves **85.3**% Answer Drift issue on average). This verifies the validity of our hypothesis 1, 2, as all of these methods are built upon them. Besides, combining the two methods leads to even greater improvements, demonstrating the necessity of optimizing from both of these perspectives.

Performance in the Model In §3 and §4, we probe information loss in two issues by tracing the information flow in models. Here, we repeat the attribution tracing experiments, comparing the differences before and after applying our method to further validate the effectiveness of our work. As we can see from Figure 8 and 16, our two methods (orange lines) increase the information flow from questions in two stages compared to CoT prompting (blue lines). This indicates our method indeed compensates for the information loss in the LLM.

Performance on Other Tasks Our work mainly focuses on commonsense reasoning tasks. To further evaluate the effectiveness of our approach on other types and forms of tasks, we conduct the main experiments on two logical reasoning tasks: **ProofWriter** (Tafjord et al., 2021) and **FOLIO** (Han et al., 2022)). As presented in Table 3, we

Method	Wino	CSQA	Hella	SIQA	PIQA	Avg
CoT	3.2k	2.6k	4.2k	2.9k	2.7k	3.1k
SC	4.5k	3.2k	5.5k	3.8k	4.3k	4.3k
SR	9.2k	6.9k	10.2k	7.6k	6.5k	8.1k
L2M	8.7k	7.9k	11.4k	8.3k	6.7k	8.6k
CON	4.6k	3.5k	5.4k	3.8k	3.5k	4.2k
Ours	3.6k	2.9k	4.9k	3.4k	3.3k	3.6k

Table 4: Token consumption per example comparison.

can find that the Toxic CoT problem also exists in other tasks, and our method can still mitigate it.

Cost Analysis For the applicability, we measure the computation and time cost of our approach. Here we compare the token cost between our method and the baseline. According to Table 4, our method requires fewer tokens compared to other SOTA methods (only $1.2\times$ cost of the basic CoT method). We also compare the time cost of our decoding method in Appendix G and find that the speed of RD is comparable to existing decoding strategies. Therefore, we illustrate the cost-efficiency of our approach across different tasks.

6 Related Work

6.1 CoT Problems Analysis and Mitigation

Recently, many works have focused on analyzing and mitigating problems in CoT reasoning. For analytical work, most studies focus on black-box LLMs. Through intervening or paraphrasing prompts and comparing outputs, researchers can interpret the reasons leading to errors in the model's reasoning (Lanham et al., 2023; Liu et al., 2023c; Wang et al., 2023a). For optimization works, they design additional supervision signals or training processes for the model (Ramnath et al., 2023; Liu et al., 2023a) or leverage external resources for the model (Shinn et al., 2023; He et al., 2023; Lyu et al., 2023). However, these works lack the probing of inner mechanisms behind these problems, leading to insufficient analysis or less universally applicable optimization methods.

6.2 Mechanistic Interpretability

The work on mechanistic interpretability aims to understand the internal mechanisms of models when performing various tasks. Early work focused on how the model stores factual knowledge internally (Meng et al., 2022; Dai et al., 2022). In recent times, some research efforts have shifted towards examining how models retrieve and utilize knowledge. This includes internal knowledge

retrieval (Geva et al., 2023), knowledge retrieval from prompts (Wang et al., 2023b; Jin et al., 2024b), and the utilization of knowledge for reasoning purposes, such as math reasoning (Stolfo et al., 2023) and multi-step reasoning (Hou et al., 2023). However, there is limited existing work that explains commonsense reasoning and CoT reasoning, which are significant contributions of this work.

7 Conclusion

In this paper, we find a problem named Toxic CoT, which results in the model's reasoning deviating from the original correct answer when utilizing CoT-like prompting. Through tracing the internal information flow of the LLM with attribution tracing and causal tracing methods, we prove that this problem is mainly caused by the model's lack of information from the question in shallow attention layers when generating rationales or answers. Based on this result, we propose the $\mathbb{RIDERS}$ method to mitigate the Toxic CoT problem from both decoding and serial position perspectives. Through extensive experiments on multiple commonsense reasoning datasets, we verify the effectiveness of our approach in mitigating Toxic CoT problems and enhancing the model's overall commonsense reasoning capabilities.

Limitations

Although our work conducts an in-depth interpretation and mitigation of the Toxic CoT problem, it has several limitations. Firstly, like former commonsense reasoning works (Liu et al., 2022, 2023a; Xie et al., 2023), our research focuses on the form of multi-choice questions. This stems from the absence of effective evaluation methods for open-ended commonsense reasoning, leading to the predominance of benchmarks in this format. This calls for advancements in benchmark-related research. Secondly, we refrain from analyzing Toxic CoT problems in more reasoning tasks such as math, primarily due to the poor performance of current moderately-sized white-box models on these tasks. For instance, Llama2-13B achieves a mere 7.2% accuracy on GSM8K (Cobbe et al., 2021) without utilizing the CoT technique. This calls for developments in model-related research. We leave these limitations as our future work to explore.

References

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7432–7439. AAAI Press.

Yew Ken Chia, Guizhen Chen, Luu Anh Tuan, Soujanya Poria, and Lidong Bing. 2023. Contrastive chain-of-thought prompting. *CoRR*, abs/2311.09277.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8493–8502. Association for Computational Linguistics.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 12216–12235. Association for Computational Linguistics.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Luke Benson, Lucy Sun, Ekaterina Zubova, Yujie Qiao, Matthew Burtell, David Peng, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Shafiq R. Joty, Alexander R. Fabbri, Wojciech Kryscinski, Xi Victoria Lin, Caiming Xiong, and Dragomir Radev. 2022. FOLIO: natural language reasoning with first-order logic. *CoRR*, abs/2209.00840.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2021. Self-attention attribution: Interpreting information interactions inside transformer. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Arti-*

ficial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021, pages 12963–12971. AAAI Press.

Hangfeng He, Hongming Zhang, and Dan Roth. 2023. Rethinking with retrieval: Faithful large language model inference. *CoRR*, abs/2301.00303.

Yifan Hou, Jiaoda Li, Yu Fei, Alessandro Stolfo, Wangchunshu Zhou, Guangtao Zeng, Antoine Bosselut, and Mrinmaya Sachan. 2023. Towards a mechanistic interpretation of multi-step reasoning capabilities of language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 4902–4919. Association for Computational Linguistics.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Qiuxia Li, and Jun Zhao. 2024a. Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 16867–16878. ELRA and ICCL.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2024b. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. *CoRR*, abs/2402.18154.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamile Lukosiute, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. 2023. Measuring faithfulness in chain-of-thought reasoning. *CoRR*, abs/2307.13702.

Linhan Li, Huaping Zhang, Chunjin Li, Haowen You, and Wenyao Cui. 2023. Evaluation on chatgpt for chinese language understanding. *Data Intell.*, 5(4):885–903.

Jiacheng Liu, Skyler Hallinan, Ximing Lu, Pengfei He, Sean Welleck, Hannaneh Hajishirzi, and Yejin Choi. 2022. Rainier: Reinforced knowledge introspector for commonsense question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 8938–8958. Association for Computational Linguistics.

Jiacheng Liu, Ramakanth Pasunuru, Hannaneh Hajishirzi, Yejin Choi, and Asli Celikyilmaz. 2023a.

Crystal: Introspective reasoners reinforced with self-feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 11557–11572. Association for Computational Linguistics.

Kang Liu, Yangqiu Song, and Jeff Z. Pan. 2024. Editorial for special issue on commonsense knowledge and reasoning: Representation, acquisition and applications. *Mach. Intell. Res.*, 21(2):215–216.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023b. Lost in the middle: How language models use long contexts. *CoRR*, abs/2307.03172.

Ziyi Liu, Isabelle Lee, Yongkang Du, Soumya Sanyal, and Jieyu Zhao. 2023c. SCORE: A framework for self-contradictory reasoning evaluation. *CoRR*, abs/2311.09603.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. *CoRR*, abs/2301.13379.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Sean Welleck, Bodhisattwa Prasad Majumder, Shashank Gupta, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. *CoRR*, abs/2303.17651.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Guanghui Qin, Yukun Feng, and Benjamin Van Durme. 2023. The NLP task effectiveness of long-range transformers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 3756–3772. Association for Computational Linguistics.

Sahana Ramnath, Brihi Joshi, Skyler Hallinan, Ximing Lu, Liunian Harold Li, Aaron Chan, Jack Hessel, Yejin Choi, and Xiang Ren. 2023. Tailoring self-rationalizers with multi-reward distillation. *arXiv preprint arXiv:2311.02805*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8732–8740. AAAI Press.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social iqa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4462–4472. Association for Computational Linguistics.

Omar Shaikh, Hongxin Zhang, William Held, Michael S. Bernstein, and Diyi Yang. 2023. On second thought, let's not think step by step! bias and toxicity in zeroshot reasoning. 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 4454–4470. Association for Computational Linguistics.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. 2023. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 7035–7052. Association for Computational Linguistics.

Tianxiang Sun, Xiaotian Zhang, Zhengfu He, Peng Li, Qinyuan Cheng, Xiangyang Liu, Hang Yan, Yunfan Shao, Qiong Tang, Shiduo Zhang, et al. 2024. Moss: An open conversational large language model. *Machine Intelligence Research*, pages 1–18.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. 2021. Proofwriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 3621–3634. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4149–4158. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and finetuned chat models. *CoRR*, abs/2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023a. Towards understanding chain-of-thought prompting: An empirical study of what matters. 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 2717–2739. Association for Computational Linguistics.

Chenhao Wang, Pengfei Cao, Jiachun Li, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Qiuxia Li, and Jun Zhao. 2024. Leros: Learning explicit reasoning on synthesized data for commonsense question answering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 10303–10315. ELRA and ICCL.

Chenhao Wang, Jiachun Li, Yubo Chen, Kang Liu, and Jun Zhao. 2022. Cn-autmic: Distilling chinese commonsense knowledge from pretrained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 9253–9265. Association for Computational Linguistics.

Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023b. Label words are anchors: An information flow perspective for understanding in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 9840–9855. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023c. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Chaojie Wen, Xudong Jia, and Tao Chen. 2023. Improving extraction of chinese open relations using pre-trained language model and knowledge enhancement. *Data Intell.*, 5(4):962–989.

Yuxi Xie, Kenji Kawaguchi, Yiran Zhao, Xu Zhao, Min-Yen Kan, Junxian He, and Qizhe Xie. 2023. Self-evaluation guided beam search for reasoning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, Juntao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. 2023. Baichuan 2: Open large-scale language models. *CoRR*, abs/2309.10305.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4791–4800. Association for Computational Linguistics.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V. Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

A   Early Statistical Experiments

In this section, we conduct early experiments on existing representative commonsense reasoning datasets to analyze the prevalence of Toxic CoT problems through statistical methods.

Datasets   We utilize five representative commonsense reasoning datasets to analyze the distribution of Toxic CoT problems. The basic information of the dataset is outlined in Table 5. It is noteworthy that, owing to the extensive size of Hellaswag's dev set (over 10,000), we extract 2,000 instances for the experiment.

Metric   We design a new metric called Toxic Rate, which measures the proportion of Toxic CoT problems among all errors. Its calculation method is shown in Equation 8.

Results   The result of our early statistical experiments is reported in Table 6. Here we use ***Llama2-13B-Chat-hf*** to present the white-box LLM and use ***GPT-3.5-turbo-1106*** to present the black-box-model. The average Toxic Rates are as high as **37.0%** and **32.8%** across the five datasets, indicating that this issue cannot be ignored and warrants further investigation.

B   Toxic Reason Statistical Experiments

In this section, we manually categorize the error types of Toxic CoT problems through statistical classification. Specifically, we sample 1,000 examples from CSQA and 1,000 examples from Winogrande, classifying the Toxic CoT problems (In all of the probing experiments in the main text, we use these samples as our probing data). The results are presented in Table 7. In the inconsistent error, the model exhibits logical inconsistency with the preceding context when generating CoT or the final answers. In the factual error, the CoT contains incorrect factual knowledge, which leads to erroneous answers. The presence of question errors reflects the subpar quality of the dataset. In such cases, questions may exhibit multiple viable answers or all options are incorrect. As for the other error, the questions trigger certain refusal-to-answer mechanisms in the model (e.g., inquiries about how to commit murder), leading to the identification of incorrect answers.

As the inconsistency error constitutes the predominant portion of all reasons, our work focuses on addressing this issue. We further categorize this error into Rationale Drift and Answer Drift based on the error occurrence (see §2.2 for their definitions).

Reason	Winogrande	CSQA
Inconsistent Error	**71(78.9%)**	**48(53.9%)**
Factual Error	16(17.8%)	17(19.1%)
Question Error	2(2.2%)	21(23.6%)
Other Error	1(1.1%)	3(3.4%)
Sum	90	89

Table 7: The classification of CoT reasoning errors

C   More Details for Reasoning Tracing

Method Implementation   We introduce the attribution score method in § 3.1. In Equation 1, we set $m = 20$ following the previous works. For $F(\cdot)$, we set it as the language modeling loss (for next-token prediction) during the CoT generation. Here, we obtain this value directly from the output of the *LlamaForCausalLM* module using the *Transformers* library. In Equation 2, we partition the step num-

	Winogrande	CSQA	HellaSwag	SIQA	PIQA
Split	dev	dev	dev	dev	dev
#Sample Num	1267	1221	2000	1954	1838
#Option Num	2	5	4	3	2

Table 5: Dataset information in this work.

	Winogrande	CSQA	HellaSwag	SIQA	PIQA	Avg
Llama2-13B	43.1	28.8	34.8	37.3	41.2	37.0
GPT-3.5	34.8	37.8	27.5	37.5	26.4	32.8

Table 6: Toxic rate on different datasets and models.

bers $|N|$ in CoT based on the occurrence of periods in the text. For models, we use `Llama2-13B-Chat` and `Baichuan2-13B-Chat`.

Attribution Tracing Experiment Figure 17 illustrates the prompt we use for generating correct CoTs from drifting cases. After the generation, we will manually filter out the wrong CoT and conduct the comparative experiment. We use a 5-shot to generate CoT and concatenate it to the question for our probing experiments. Figure 9 shows the remaining results of this experiment, from which we can get the same conclusions as Section 3.2. Additionally, we also conduct the supplementary experiments on CSQA and report the results in Figure 10.

Attention Tracing Experiment Figure 11 reports more results in this experiment.

D More Details for Answering Tracing

Intervention Tracing Experiment Figure 12, 13, 14 and 15 report the remaining intervention tracing experiment results, which are consistent with our conclusion in the main text.

E Mitigation Method Implementation

Residual Decoding Here, we provide a detailed explanation of Algorithm 1. At the beginning, we set the input to the entire question (contexts + options). In line 2, we get the logits from the output of the *LlamaForCausalLM*. In line 6, we calculate the attention score by summing the values on the attention matrix corresponding to the tokens. We use the output character "*</s>*" as the termination condition for Llama2-13B generation in line 12.

Serial-Position Swap In this method, we swap the positions of the question and the generated CoT, outputting the option with the highest logits score. This method can be implemented under both few-shot and zero-shot settings, demonstrating its cost-efficiency.

F More Details for the Mitigation Experiment

Dataset In this experiment, the specific information of all datasets can be found in Table 5.

Baselines For the Self-Consistency method, we sample 5 CoTs and use a majority voting method to select the final predicted answer. For the Self-Refine method, we first conduct one round

of CoT reasoning and then follow it with one round of feedback to generate the final answer. For all the baselines, we release the prompts in our source code. We implement all of the methods on `Llama2-13B-Chat-hf` and `Baichuan2-13B-Chat`.

Our methods In our RD method, we set two hyperparameters — candidate_num n and weight ω, and here are their specific values in the experiments: for Winogrande, we set n to 4 and ω to 80, for CSQA, we set n to 10 and ω to 135, for HellaSwag, we set n to 3 and ω to 80, for SIQA, we set n to 10 and ω to 160, for PIQA, we set n to 4 and ω to 120. Additionally, in Figure 18,19,20,21,22, we list our method's few-shot prompts on five datasets. Note that both CoT prompting and our two methods utilize the same prompt.

Results In our paper, we choose Llama2-13B for experiments because it is an open-source LLM of suitable size and has great influence. Here, we also repeat all experiments on Baichuan2-13B to verify the generality of our work (see Table 8). Additionally, in Table 9, we show our experimental results on two other representative models: `Mistral-7B-Instruct-v0.2` and `gpt-3.5-turbo-1106` (since it is a closed source model, here we only apply the SPS method).

G More Details for Cost Analysis

Table 10 illustrates the time cost comparison of our residual decoding methods with other strategies. Here we set *num_beams* in the beam search strategy as 5, and *candidate_num* in the RD strategy as 3. We compute the average seconds cost per example over 50 samples for each dataset. On average, our decoding strategy takes 2.6 times longer than greedy search and 1.4 times longer than beam search. This reflects that our decoding method has significantly stronger performance while having a comparable overall time to these main decoding strategies.

Method		Winograffe		CSQA		HellaSwag		SIQA		PIQA		Avg	
		ACC↑	TR↓	ACC↑	TR↓	ACC↑	TR↓	ACC↑	TR↓	ACC↑	TR↓	ACC↑	TR↓
Few-shot Answer		59.7	-	69.9	-	**47.1**	-	68.8	-	60.7	-	61.2	-
Chain-of-Thought		58.7	34.8	66.8	37.8	41.1	27.5	66.5	37.5	66.1	26.4	59.8	32.8
Self-Consistency		56.1	38.8	64.3	41.5	39.4	30.6	67.8	33.5	**70.0**	21.0	59.5	33.1
Self-Refine		59.6	35.7	66.7	38.8	39.2	29.9	62.2	43.1	62.6	31.0	58.1	35.7
Least-to-Most		59.7	35.6	62.8	45.6	37.8	32.1	66.5	36.8	62.8	20.6	57.9	34.1
Contrasive CoT		55.2	40.6	67.0	37.4	38.7	29.7	65.0	37.3	65.9	26.2	58.4	34.2
Ours	RD Only	59.2	21.1	69.5	25.3	44.4	19.3	68.9	25.0	67.0	12.9	61.8	20.7
	SPS Only	59.9	22.8	69.7	23.5	46.3	15.2	**69.2**	27.8	67.7	14.2	62.6	20.7
	RIDERS	**60.1**	**18.6**	**71.3**	**14.6**	46.3	**13.0**	69.0	**20.5**	67.2	**8.3**	**62.8**	**15.0**

Table 8: Performance comparison across five commonsense reasoning datasets on Baichuan2-13B.

Model	Method	Winograffe		SIQA	
		ACC↑	TR↓	ACC↑	TR↓
Mistral-7B	CoT	59.2	27.9	70.2	28.2
	RIDERS	**62.0**	**15.3**	**70.6**	**24.5**
GPT-3.5	CoT	64.6	48.8	70.6	30.3
	RIDERS	**70.1**	**41.7**	**72.1**	**29.0**

Table 9: Performance comparison on Mistral-7B and GPT-3.5.

Strategy	Wino	CSQA	Hella	SIQA	PIQA	Avg
Greedy	5.2	2.8	4.9	3.6	5.8	4.5
Beam	8.1	4.9	10.1	5.8	9.4	7.7
Ours	13.4	6.6	16.2	7.4	15.9	11.9

Table 10: Decoding time per example comparison.

(a) Llama2-13B
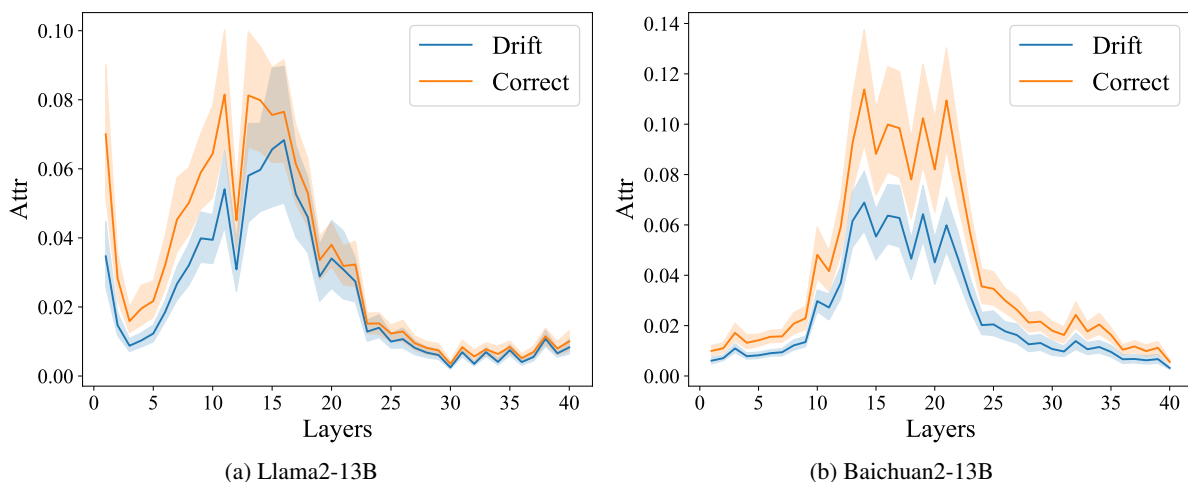
(b) Baichuan2-13B

Figure 9: Attribution tracing results on CSQA.

(a) Llama2-13B
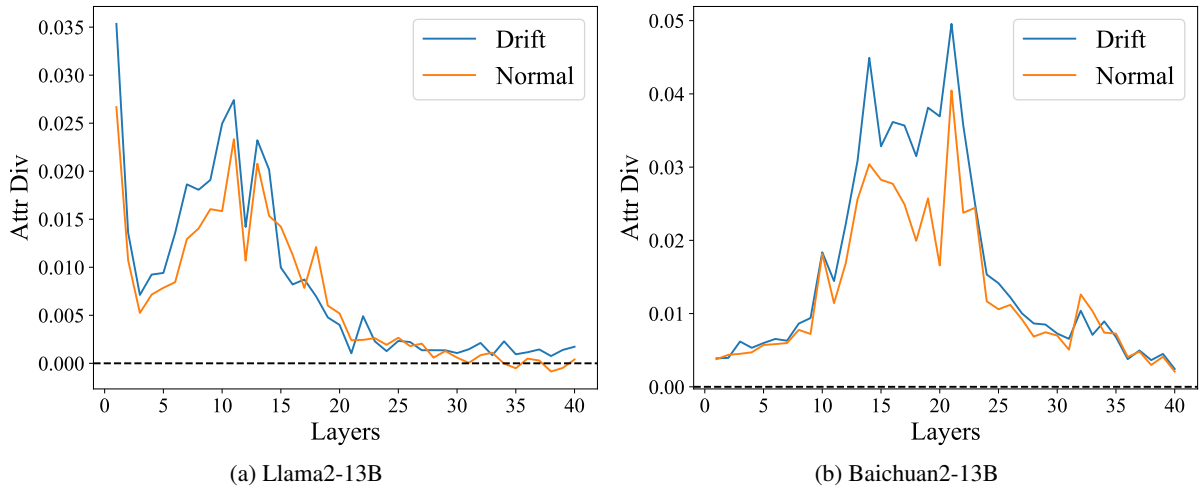
(b) Baichuan2-13B

Figure 10: Information flow divergence comparison on CSQA.
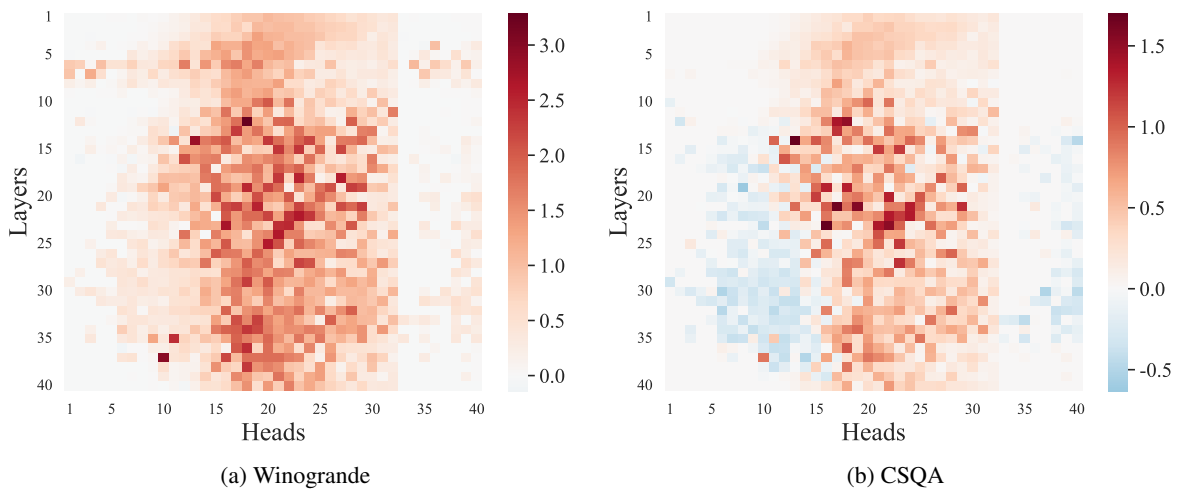

(a) Winogrande

(b) CSQA

Figure 11: Attention tracing results across different attention heads on Baichuan2-13B.
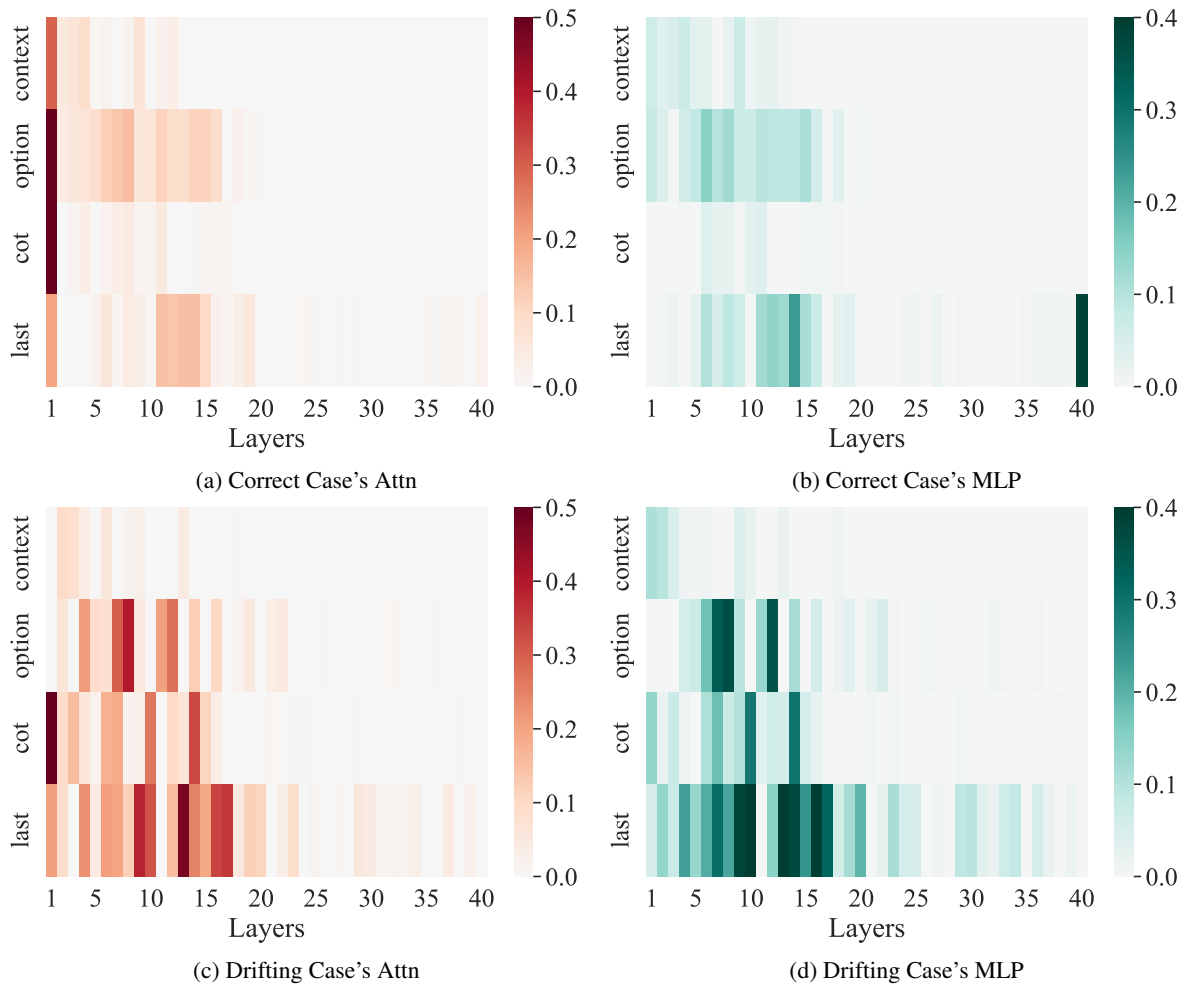
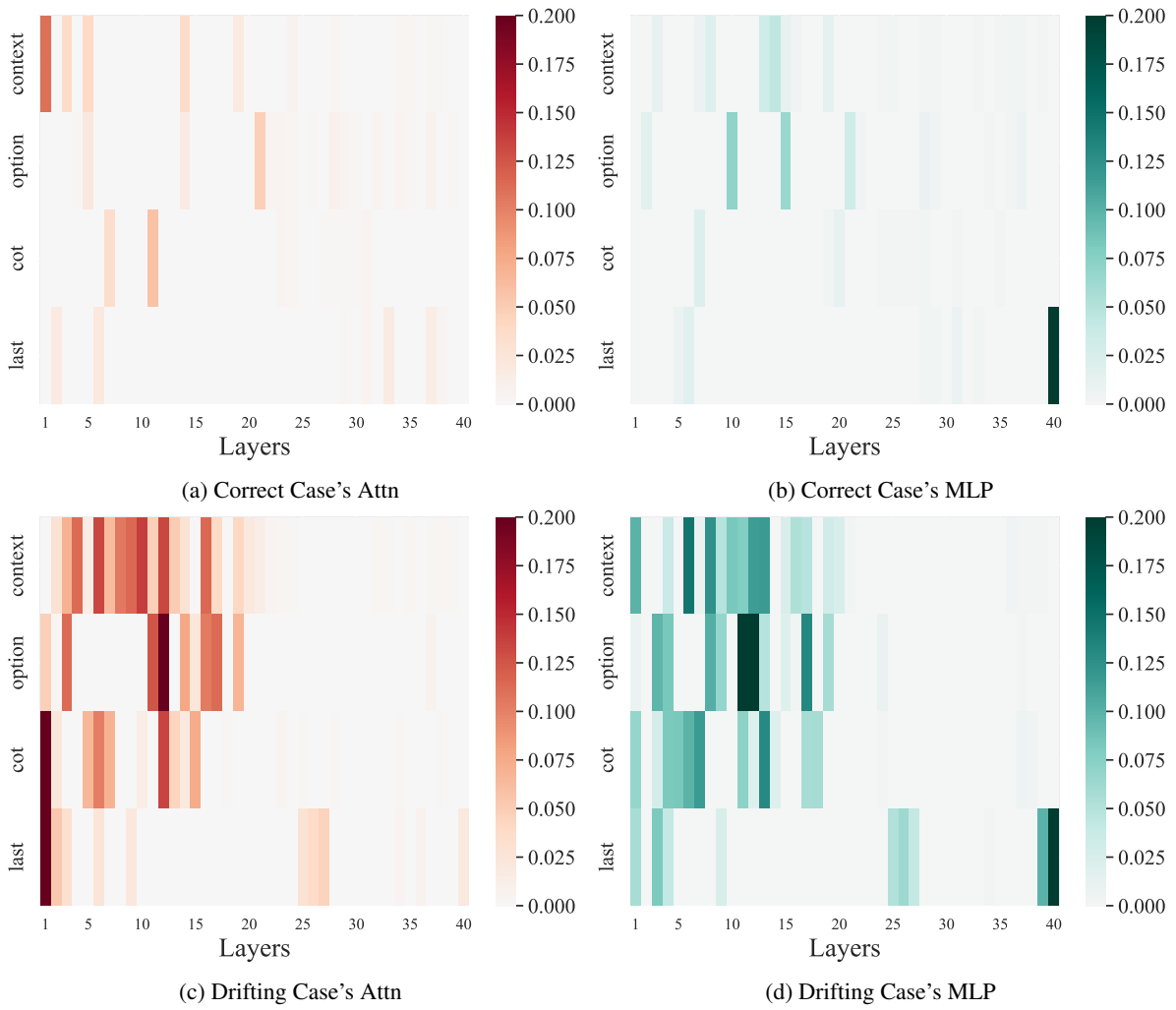Figure 12: Intervention tracing results on CSQA in correct and drifting answering cases (Llama2-13B).

(a) Correct Case's Attn

(b) Correct Case's MLP

(c) Drifting Case's Attn

(d) Drifting Case's MLP

Figure 13: Intervention tracing results on Winogrande in correct and drifting answering cases (Baichuan2-13B).

(a) Correct Case's Attn
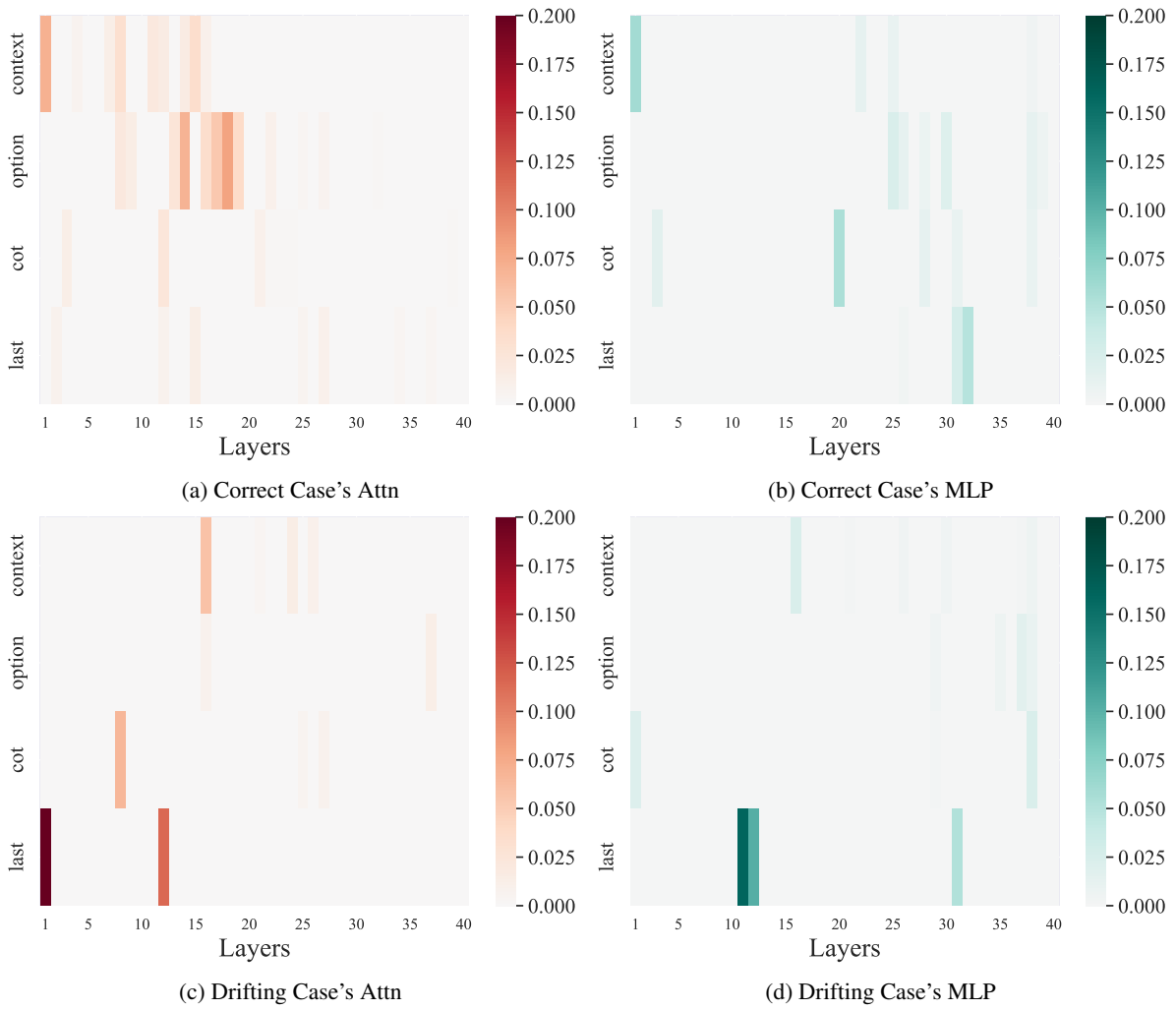
(b) Correct Case's MLP

(c) Drifting Case's Attn

(d) Drifting Case's MLP

Figure 14: Intervention tracing results on CSQA in correct and drifting answering cases (Baichuan2-13B).

(a) Winogrande

(b) CSQA
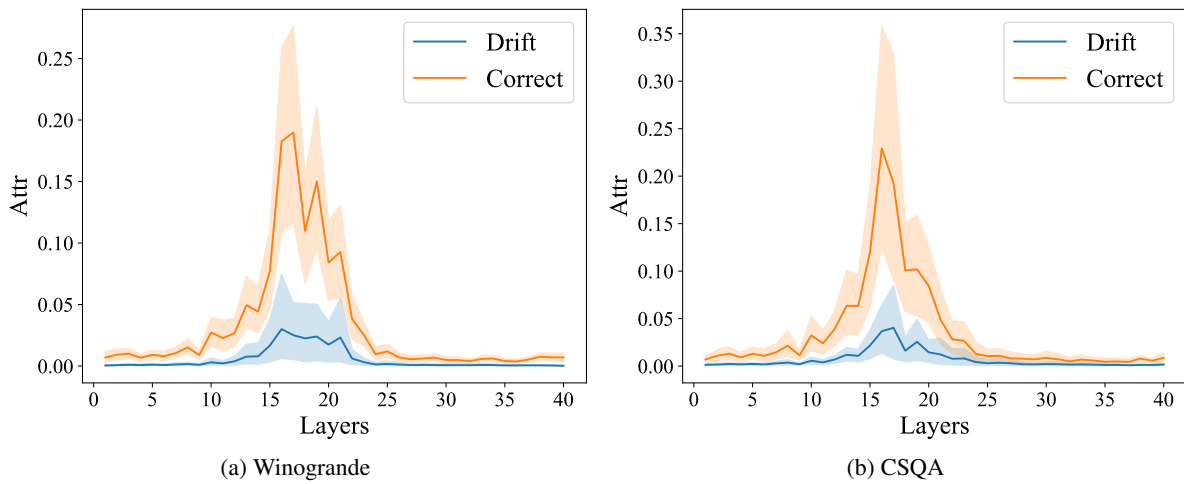
Figure 15: Attribution tracing results on Baichuan2-13B during the answer generation stage.
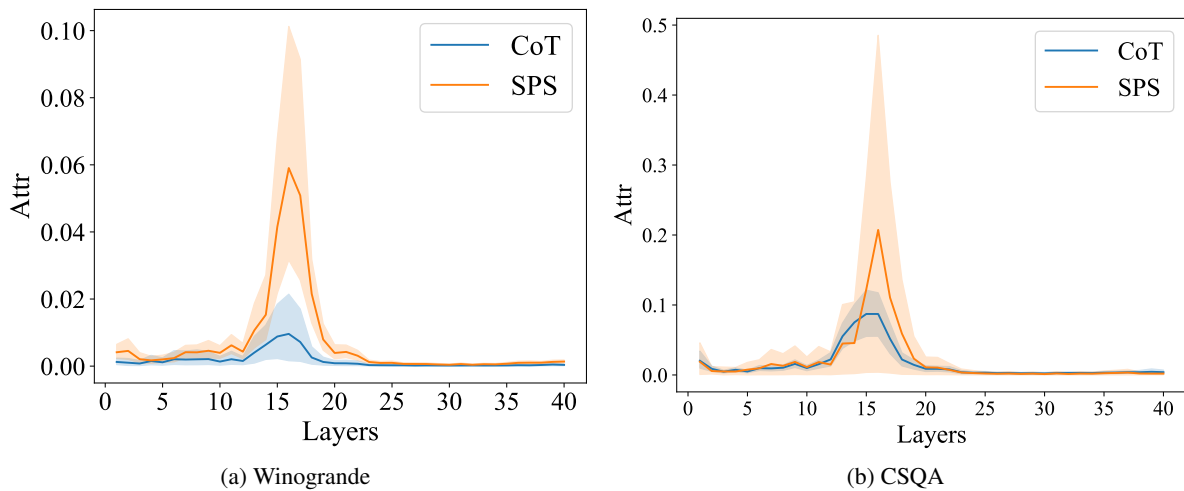

(a) Winogrande

(b) CSQA

Figure 16: Information flow comparison on CSQA after applying our two methods.

System Instruction: You are a helpful, respectful and honest assistant. You should use your reasoning abilities to given the reason for the answer to the given questions. Your response should be in this form: 'Reason: {reason}'.

Example Input: Question: {QUESTION}\nAnswer: {ANSWER}
Example Output: Reason: {COT}

Input: Question: {QUESTION}\nAnswer: {ANSWER}
Output:

Figure 17: Prompts for correct CoT generation.

System Instruction: You are a helpful, respectful and honest assistant. You should use your reasoning abilities to answer the given questions in reasoning tasks. You should reply the correct rationales and the answer. Your response should be in this form: '{reason} So the answer is: ({option}) {answer}'. If you don't know the answer to a question, please reply 'Answer: None'.

Question: The test was hard for Samuel but a breeze for Randy , since _ had failed to study for it.\n(1) Samuel (2) Randy
Answer: To pass a test, a person need to study for it. If a person feel the test like a breeze, it means the test is easy for him. A person feels the test easy, because he studies hard for it. Since we know that Samuel feel the test very hard, she may fail to study for it. So the answer is: (1) Samuel.

Question: Kyle slowly wormed their way into the life of Derrick, because _ was good and manipulating people.\n(1) Kyle (2) Derrick
Answer: A person wormes his way into other's life, because he is friendly and approachable. A friendly person is considered good. A person is seen as manipulating people, that means he like to interact with others and others like him. Since Kyle slowly wormed their way into the life of Derric, Kyle will be seen as good and manipulating people. So the answer is: (1) Kyle.

Question: Donald was very grounded but Michael often got lost in their daydreams. _ was very capricious all the time.\n(1) Donald (2) Michael
Answer: A person is grounded means he works hard and does not like to fantasize. A person often gets lost in his daydreams, he is seen as unrealistic and egocentric. A person is capricious all the time means he does everything only according to his own ideas. Since Michael often gets lost in their daydreams but Donald does not, Michael is seen as very capricious all the time. So the answer is: (2) Michael.

Question: After stopping when running, Betty was able to help Rachel treat their asthma because _ has it too.\n(1) Betty (2) Rachel
Answer: A person who has a mild disease may knows how to treat it. Since Since Betty was able to help Rachel treat their asthma, Betty may has it, too. So the answer is: (1) Betty.

Question: Jeffrey was sick with the stomach flu and not hungry, but Christopher was starving. _ ordered food.\n(1) Jeffrey (2) Christopher
Answer: Stomach flu is a viral infection in the digestive system. A person is starve so that he need to eat something next. A person is not hungry means he does not want to eat something. Ordering food means to buy food online. A person may eat food after ordering it. Since Christopher was starving and Jeffery was sick and not hungry, Christopher tended to order food. So the answer is: (2) Christopher.

Figure 18: 5-shot prompts for Winogrande.

System Instruction: You are a helpful, respectful and honest assistant. You should use your reasoning abilities to answer the given questions in reasoning tasks. You should reply the correct rationales and the answer. Your response should be in this form: '{reason} So the answer is: ({option}) {answer}'. If you don't know the answer to a question, please reply 'Answer: None'.

Question: Google Maps and other highway and street GPS services have replaced what? \n (1) atlas (2) mexico (3) countryside (4) united states (5) oceans **Answer:** Electronic maps and GPS services are the modern version of paper atlas. In that case, the atlas have been replaced by Google Maps and other highway and street GPS services. So the answer is: (1) atlas.

Question: The fox walked from the city into the forest, what was it looking for? \n (1) pretty flowers. (2) hen house (3) natural habitat (4) storybook (5) dense forest
Answer: Since the fox walk from the city into the forest, he may looks for something in the forest but not in the city. From all of the options, the natural habitat are usually away from cities. So the answer is: (3) natural habitat.

Question: You can share files with someone if you have a connection to a what? \n (1) freeway (2) radio (3) wires (4) computer network (5) electrical circuit
Answer: Files usually can be stored in the computers. In that case, we can share them over the Internet. Thus, if we connect to a computer network, we can share the file with others. So the answer is: (4) computer network.

Question: Too many people want exotic snakes. The demand is driving what to carry them? \n (1) ditch (2) shop (3) north america (4) outdoors (5) pet shops
Answer: If people want exotic snakes, they may like to raise snakes as pets. If there is a demand for snakes as pets, pet shops will be pushed to carry them, in order to make more money. So the answer is: (5) pet shops.

Question: The body guard was good at his duties, he made the person who hired him what? \n (1) better job (2) feel safe (3) irritated (4) save money (5) headache
Answer: The job of body guards is to ensure the safety and security of the employer. People ususally hire the body guard to make themselves safe. So the answer is: (2) feel safe.

Figure 19: 5-shot prompts for CSQA.

System Instruction: You are a helpful, respectful and honest assistant. You should use your reasoning abilities to answer the given questions in reasoning tasks. You should reply the correct rationales and the answer. Your response should be in this form: '{reason} So the answer is: ({option}) {answer}'. If you don't know the answer to a question, please reply 'Answer: None'.

Question: Then, the man writes over the snow covering the window of a car, and a woman wearing winter clothes smiles. then \n (1) the man adds wax to the windshield and cuts it. (2) a person board a ski lift, while two men supporting the head of the person wearing winter clothes snow as the we girls sled. (3) the man puts on a christmas coat, knitted with netting. (4) the man continues removing the snow on his car.
Answer: Snow must be removed from a car before one can drive it. Since there is some snow covering the car, so the man tends to remove it next. So the answer is: (4) the man continues removing the snow on his car.

Question: The man in the center is demonstrating a hairstyle on the person wearing the blue shirt. the man in the blue shirt \n (1) is standing on the sponge cutting the hair of the person wearing the blue shirt. (2) is doing the hairstyle with his hand and the hairspray. (3) sits on the chair next to the sink. (4) is being shown eye to eye.
Answer: One should sit still when getting a haircut. People often sits down when they are get a haircut. From the context, we can get that the man in the blue shirt is getting a haircut, so he should sit down and wait. So the answer is: (3) sits on the chair next to the sink.

Question: The roof is done and a view of the entire house is shown to show off the finished roof. the woman \n (1) is standing in front of the home, smiling while talking. (2) interviews the man again and leaves the room. (3) shows the soil with two lay-ups of shingle and applies a layer onto the top shingle. (4) stacks the bags on the side and begins putting stencils on the top.
Answer: One usually feels pleased after finishing a home improvement project. If a person feels pleased, he tends to smile. Since the woman has finished all the work, she may be proud of it and feel pleased. Then she may smile. So the answer is: (1) is standing in front of the home, smiling while talking.

Question: People practice ballet in a studio alone and in couples. then \n (1) a man begins dancing and top dancing standing on the ground. (2) a boy and a girl dance ballet, then a man enter and dance with the girl. (3) the band performs ballet in the studio and in the open gathered. (4) people dances and dances together, dancing.
Answer: Ballet is a kind of dance. If people pratice ballet, they often dance with their partener or dance alone, which matches the description of option (2). So the answer is: (2) a boy and a girl dance ballet, then a man enter and dance with the girl.

Question: A person is seen standing on a tennis court bouncing a ball. another man \n (1) takes his turn speaking to the camera. (2) walks up hitting a birdie. (3) is seen standing ready on the other side in front of a large audience. (4) is holding the racket next to him.
Answer: The player who bounces the ball is the one who serves. In the tennis race, two players are located on opposite sides of the court, one is serving and the other is preparing to receive the ball. Since there is a person who serves, another one should stands ready. So the answer is: (3) is seen standing ready on the other side in front of a large audience.

Figure 20: 5-shot prompts for HellaSwag.

System Instruction: You are a helpful, respectful and honest assistant. You should use your reasoning abilities to answer the given questions in reasoning tasks. You should reply the correct rationales and the answer. Your response should be in this form: '{reason} So the answer is: ({option}) {answer}'. If you don't know the answer to a question, please reply 'Answer: None'.

Question: Quinn wanted to help me clean my room up because it was so messy. What will Quinn want to do next? \n (1) Eat messy snacks (2) help out a friend (3) Pick up the dirty clothes
Answer: Quinn want to clean the room up. Picking up the dirty clothes is one way to clean the room. Thus, quinn will want to pick up the dirty clothes next. So the answer is: (3) Pick up the dirty clothes

Question: Sasha's mom passed out in the middle of the party. Aubrey took Sasha's mom to the hospital. What will Aubrey want to do next? \n (1) help Aubrey go back home (2) keep on partying without the mom (3) going on with the mom
Answer: Sasha's mom is sent to the hospital beacuse she passes out. After that, she becomes a patient. A patient needs to be cared for. Aubrey may need to take care of her. Thus, Aubrey will go on with the mom. So the answer is: (3) going on with the mom

Question: Their cat kept trying to escape out of the window, so Jan placed an obstacle in the way. How would Jan feel afterwards? \n (1) scared of losing the cat (2) normal (3) relieved for fixing the problem
Answer: The cat try to escape so Jan need to stop it to avoid losing the cat. Jan place an obstacle in the way so the cat can not escape. The problem is been solved. Thus, Jan will fell relieved for fixing the problem. So the answer is: (3) relieved for fixing the problem

Question: Sydney had so much pent up emotion, they burst into tears at work. How would Sydney feel afterwards? \n (1) affected (2) like they released their tension (3) worse
Answer: Crying is often a way to release tension. Sydney burst into tears at work. Thus, she would release the tension. So the answer is: (2) like they released their tension

Question: Sydney got a raise and a new promotion. What does Sydney need to do before this? \n (1) be bad at her job (2) do a good job (3) be lazy
Answer: To get a raise and a new promotion, Sydney needs to work hard and do a good job. Sydney got a raise and a new promotion. Thus, she need to do a good job before this. So the answer is: (2) do a good job

Figure 21: 5-shot prompts for SIQA.

System Instruction: You are a helpful, respectful and honest assistant. You should use your reasoning abilities to answer the given questions in reasoning tasks. You should reply the correct rationales and the answer. Your response should be in this form: '{reason} So the answer is: ({option}) {answer}'. If you don't know the answer to a question, please reply 'Answer: None'.

Question: How do you flood a room?\n(1) fill it with objects. (2) fill it with water
Answer: Too much water can cause flooding. Thus, if we want to flood a room, we should use water. So the answer is: (2) fill it with water

Question: How can I get oil stains out of my driveway?\n(1) Douse each stain with a couple cans of beer. (2) Douse each stain with a couple cans of soda.
Answer: Sodium carbonate solution can wash away oil stains. The soda is a kind of sodium carbonate solution. Thus, you can use cans of soda to get oil stains out of your driveway. So the answer is: (2) Douse each stain with a couple cans of soda.

Question: Soothe a painful sunburn.\n(1) Wait until brewed tea bag is cool, then apply on burn. (2) Wait until brewed tea bag is hot, then apply on burn.
Answer: Sunburn can be alleviated by applying cold material. Thus, you should apply cool tea rather than hot tea bag to soothe your sunburn. So the answer is: (1) Wait until brewed tea bag is cool, then apply on burn.

Question: What can I use for fuel in an alcohol stove?\n(1) Use acetone. (2) Use vinegar.
Answer: Acetone is flammable, while vinegar is not. If you want to use something for fuel, the thing you use should be flammable. Thus, you should use acetone for fuel in an alcohol stove. So the answer is: (1) Use acetone.

Question: How can I cut the handles of metal cutlery?\n(1) Use a hand saw to cut the handles. (2) Use a hand drill to cut the handles.
Answer: A hand saw is used for making cuts and a hand drill is used for making holes. If you want to cut something, you should use a hand saw rather than hand drill. So the answer is: (1) Use a hand saw to cut the handles.

Figure 22: 5-shot prompts for PIQA.