Making Reasoning Matter: Measuring and Improving Faithfulness of Chain-of-Thought Reasoning

Anonymous ACL submission

Abstract

Large language models (LLMs) have been shown to perform better when asked to reason step-by-step before answering a question. However, it is unclear to what degree the model's final answer is faithful to the stated reasoning steps. In this paper, we perform a causal mediation analysis on twelve LLMs to examine how intermediate reasoning steps generated by the LLM influence the final outcome and find that LLMs do not reliably use their intermediate reasoning steps when generating an answer. To address this issue, we introduce FRODO, a framework to tailor small-sized LMs to generate correct reasoning steps and robustly reason over these steps. FRODO consists of an *inference module* that learns to generate correct reasoning steps using an implicit causal reward function and a *reasoning module* that learns to faithfully reason over these intermediate inferences using a counterfactual and causal preference objective. Our experiments show that FRODO significantly outperforms four competitive baselines. Furthermore, FRODO improves the robustness and generalization ability of the reasoning LM, yielding higher performance on out-of-distribution test sets. Finally, we find that FRODO's rationales are more faithful to its final answer predictions than standard supervised fine-tuning.

1 Introduction

Chain-of-thought (CoT) reasoning techniques have been shown to improve the performance of large language models (LLMs) by generating step-by-step reasoning traces before generating a final answer (Wei et al., 2022). Many works suggest that the reasoning process described in CoT explanations may be a possible description of how models make predictions (Kojima et al., 2022; Yao et al., 2023; Sun et al., 2023). However, despite the remarkable success of CoT in many reasoning tasks, recent works show that LLMs-generated reasoning traces can be incorrect (Zhang et al., 2023) and unfaithful (Turpin et al., 2023).

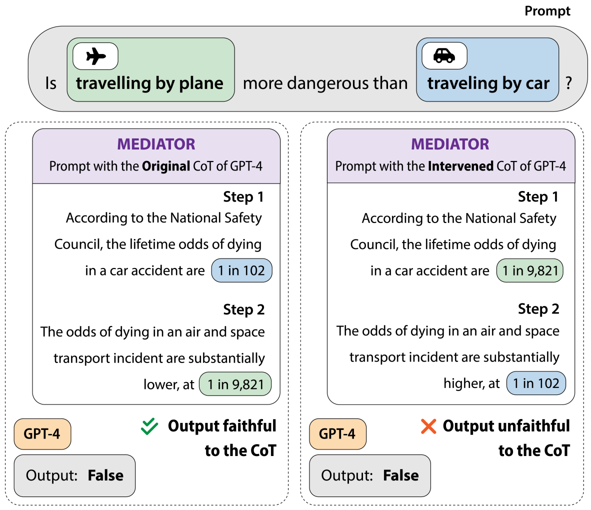

Figure 1: An example of our proposed causal analysis to measure the faithfulness of the final output to the CoT generated by the model. We perturbed CoT rationales and studied the causal impact on the model's behaviour.

Reasoning implicitly involves two steps: identifying the rules and facts (inference chains) necessary to reach a conclusion and then robustly using them to reach said conclusion (Levesque, 1986). Our paper studies whether LLMs reliably use inference chains to arrive at a conclusion.[1] In standard CoT, LLMs can generate plausible explanations with the final answer not necessarily guaranteed to follow the reasoning chain or imply a causal relation between the reasoning chain and the model's outcome (Lyu et al., 2023). Most recent efforts have either focused on the performance of LLMs on various reasoning tasks or their faithfulness in CoT generation, ignoring the sequential relationship between CoT and the final answer (Huang and Chang, 2023; Lanham et al., 2023).

[1]In our paper, reasoning faithfulness refers to how reliably the model uses its reasoning steps to arrive at a correct answer.

In this work, we address this gap by introducing a methodology for interpreting the relationship between the CoT trace and the final answer based on causal mediation analysis (Pearl, 2001). Causal mediation analysis is a method of causal inference that studies the change in a response variable following an intervention or treatment. More concretely, we use this method to measure and interpret the contribution of a reasoning chain (mediator) to the final answer (observed output), as shown in Fig.8. We propose multiple interventions on the model inputs and mediators (reasoning chain) to unveil the causal effect of specific reasoning steps in a model's output.

We apply this framework and study the causal impact of CoT rationales on the behaviour of twelve different state-of-the-art LLMs on three different complex reasoning tasks (*mathematical, commonsense*, and *causal understanding*). We observe a large variation across tasks and models in how strongly reasoning traces causally affect the model's prediction. In particular, we find that instruction-tuned models (GPT-3.5-Instruct, Brown et al., 2020b; Mistral-Instruct-7B, Jiang et al., 2023) have a stronger causal effect on the final answer when conditioned on the reasoning trace than models trained with RLHF (e.g., ChatGPT; Llama-2-7B-Chat, Touvron et al., 2023). Similar to Turpin et al. (2023), when we intervene in the reasoning problem, we observe that ChatGPT and GPT-3.5-Instruct are inconsistent at generating plausible reasoning chains. Finally, we find GPT-4 (Achiam et al., 2023) only changes its answer 30% of the time when conditioned on perturbed counterfactual reasoning chains. In Figure 8, we see one example where GPT-4 does not faithfully change its final answer when provided with intervened counterfactual CoT. These results indicate two issues: (i) LLMs can generate unfaithful and implausible reasoning chains, and (ii) LLMs are inconsistent when reasoning over their own generated reasoning traces.

To address these issues, we introduce a novel method, FRODO, comprising two modules. The first module tailors small-sized LMs to generate correct reasoning chains (inference module), while the second module takes the reasoning chains as input and faithfully reasons over them to arrive at the correct answer (reasoning module). To learn to generate correct reasoning chains, we use the DPO algorithm (Rafailov et al., 2023), which enables the model to prefer correct reasoning chains over counterfactual ones with implicit feedback. Instead of relying on human labeling, we obtain preference data by prompting LLMs to generate correct and counterfactual reasoning chains. Second, we train another small-sized LM to improve the causal effect between the reasoning chain and the final answer using a counterfactual and causal preference ranking objective.

We evaluate FRODO on four reasoning tasks (Quarel, StrategyQA, OpenBookQA, QASC) using multiple model backbones of different scales, and demonstrate that FRODO achieves an absolute accuracy improvement of $2\% \sim 3\%$ over standard supervised fine-tuning or CoT distillation methods. We assess robustness by examining how models alter their answers when intervened with counterfactual reasoning chains. FRODO exhibits significant (+4.5%) improvement in robustness. Finally, FRODO generalizes better to out-of-distribution test sets, showing a +2.6% performance improvement over supervised fine-tuning.[2]

2 Reasoning Chain as a Mediator

2.1 Problem Formulation

Reasoning is often a process that involves composing multiple inference steps to reach a conclusion or make a decision. We informally conceptualize each reasoning task as requiring a model $f\colon X \rightarrow Y$ to map an input $x \in X$ to an output $y \in Y$ by making correct or plausible inference steps R.

Causal Interpretation. The causal graph is a probabilistic graphical model used to describe how variables interact, expressed by a directed acyclic graph consisting of the sets of nodes (N) denoting the variables and the edges (E) between the nodes denoting the causality (Pearl, 1998). As shown in Fig. 2(a), if Y is a descendent of X, then X is a potential cause of Y, and Y is the effect.

Causal mediation analysis. It is a method of causal inference which studies the change in a dependent variable following an intervention or treatment of an independent variable (Pearl, 2001). Causal mediation analysis aims to decompose the total effect of the independent variable (X) on the dependent variable (Y) into two components: the direct effect and the indirect effect (Pearl, 2001). In this work, we view the reasoning process as a causal graph, framing the input (reasoning problem) X and the output Y as random variables and the reasoning steps as mediator variable R. We use

[2] Code and data will be released upon publication.

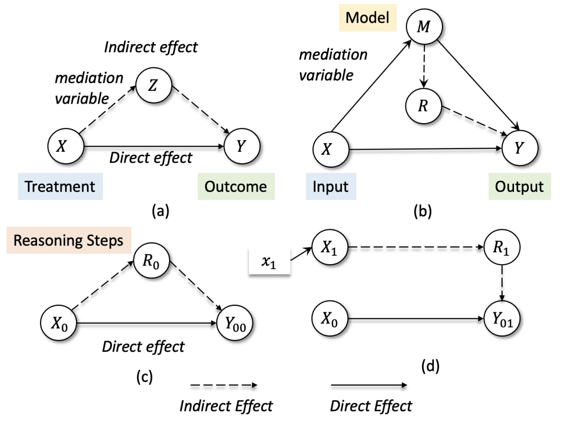

Figure 2: Causal graph for natural language reasoning, where we model $P(Y|do(x))$. X_0 = original reasoning problem, X_1 = intervened reasoning problem, R_0 = reasoning steps for X_0 reasoning problem, R_1 = reasoning steps for X_1, Y = output, and M = model parameters

Variables	Example
X_0	*Is Poseidon similar to the god Vulcan?*
X_1	*Is Poseidon similar to the god Neptune?*
R_0	*Poseidon is a god from Greek mythology, known as the god of the sea, earthquakes, and horses. Vulcan is a god from Roman mythology, known as the god of fire, metalworking, and the forge. Although both are gods, they represent different elements and aspects, and come from different mythologies.*
R_1	*Poseidon is a god from Greek mythology, known as the god of the sea, storms, and earthquakes. Neptune is a god from Roman mythology, who is also known as the god of the sea. Both Poseidon and Neptune share similar roles and attributes in their respective mythologies.*

Table 1: An example from StrategyQA dataset, where X_1, R_0 and R_1 are generated by GPT-4.

mediation analysis to interpret the role of reasoning steps as mediators between model inputs and model outputs. Please note that there are two mediator variables, model parameters M and reasoning steps R (see Fig. 2(b)). However, in this work, we are interested in understanding the indirect effect of R on the output Y; hence, we assume model parameters M to be fixed. Let X_0 denote the initial reasoning problem, R_0 as the reasoning chain given X_0 and Y_{00} denote the potential outcome if the treatment variable and mediator variable are X_0 and R_0, respectively. Whereas Y_{01} denotes the possible outcome when treatment is set to X_0 and R_1 is the reasoning chain for an intervened reasoning problem (see Fig.2(d)).

Direct Effect (DE) measures how much an intervention X changes an outcome variable Y directly, without passing through a hypothesized mediator R. Based on Fig 2(c), we define the direct effect of X=X_0 on Y as $DE = P(Y|X_0,R_0) - P(Y|X_1,R_0)$ which can be seen as the comparisons between two potential outcomes of X given two different treatments, i.e., $X = X_0$ and X_1. It is computed by applying the intervention X but holding R fixed to its original value (R_0).

Indirect Effect measures how much the intervention X changes Y indirectly through R. We define the indirect effect as $IE = P(Y|X_0,R_0) - P(Y|X_0,R_1)$. It is computed by setting R to its value under the intervention X while keeping everything else to its original value.

More concretely, according to Pearl (2001), in our scenario, a high direct effect means that X has higher causality on Y, i.e., the model puts more emphasis on the reasoning problem than the reasoning steps. In contrast, a high indirect effect means the model puts more emphasis on the reasoning steps than the problem input.

Reasoning Intervention. Following Pearl (2001), we conduct counterfactual reasoning to measure the causal effect of a treatment variable on a response variable. We first perform targeted interventions on the input text X and measure their effect on the reasoning outcome Y (direct effect). Further, we also perform interventions on the mediator R and measure their effect on Y (indirect effect). We perform the following steps to automatically generate an intervention on X and R.

Step 1: Intervention Data Generation. We use a large language model (GPT-4) to automatically generate an alternative treatment variable X_1[3]. The input to M includes instruction and few-shot examples, taking the format shown in Table 14. LLMs can be sensitive to instructions and few-shot examples; hence, we randomize the prompt by manually creating a set of semantically similar instructions. Then, we randomly sample from the instruction set each time.

Step 2: Manual Data Curation. To retain high-quality data for our analysis, we manually filter out generated samples from Step 1 that are invalid or low-quality. Table 1 shows an example where given the original input reasoning question X_0, the model generated X_1, where it replaces "*Vulcan*" with "*Neptune*". To study the direct effect by taking the difference of the model output when we provide them X_0, R_0 and X_1, R_0 as inputs.

Step 3: Generate Reasoning Chain. Finally, to get the indirect effect, we generate the reasoning chain (R_0, R_1) for each reasoning problem X_0

[3] See App. A.6 for details on task-specific interventions.

Models	StrategyQA			GSM8k			Causal Understanding		
	CoT (%)	NIE	NDE	CoT (%)	NIE	NDE	CoT (%)	NIE	NDE
ChatGPT	69.2	15.3	9.1	70.1	56.3	1.01	58.8	21.1	27.4
GPT-4	93.5	40.0	22.2	81.1	21.01	30.01	72.5	29.1	48

Table 2: **Causal Effects** of generated CoT and reasoning problems on the outputs, with both Natural Indirect Effect (NIE) and Natural Direct Effect (NDE). COT (%) represents the accuracy of the models.

or X_1 by providing LLMs with some high-level descriptions about each reasoning task and reasoning prompt – "*Let's think step by step*"(see App. A.5).

2.2 Causal Mediation Analysis Results

In Table. 2, 3, we report the results of the causal mediation analysis for twelve models. In section §4, we provide the details about the implementation, evaluation metrics and datasets.

Natural direct and indirect effects. We first evaluate the indirect and direct effects of the reasoning chain and reasoning problems on the final outputs. For models (>100B) with the emergent ability to generate plausible reasoning chains, we report *natural* indirect effects and direct effects (see §2). Table 2 shows the zero-shot performance of the ChatGPT and GPT-4 models. We observe that for StrategyQA and Causal Understanding tasks, GPT-4 has a higher natural indirect effect than ChatGPT, suggesting that it is able to better reason over the reasoning steps for these tasks. However, for mathematical reasoning (GSM8K), ChatGPT has a better indirect effect. Qualitatively, we find that for mathematical reasoning, when we provide intervened reasoning steps, GPT-4 considers them incorrect and continues to generate correct reasoning steps. This results in a lower indirect effect score. Moreover, GPT-4 exhibits a more pronounced direct effect than ChatGPT, suggesting that its outputs are more causally sensitive to reasoning problems. In general, our experiments show a large variation in the causal effects of COT in the final answer depending on the tasks.

Controlled direct and indirect effects. Table 3 shows the results of causal mediation analysis for 12 different LMs. In these experiments, we examined the causal behaviour using reasoning chains generated by GPT-4 (controlled setting). Our study suggests that vanilla LMs (<20B) (in a zero-shot setting) are systematically unfaithful and consistently fail to reason over the mediator. We find that in-context learning and instruction-tuning improve the indirect effect over models trained only with language modelling objectives (e.g., LLaMA and Mistral), indicating that these methods help the model align better with the reasoning chains. We observe that models trained with RLHF objective (ChatGPT, Llama-2-7B-Chat) have a more direct effect than an indirect effect, suggesting that training on human feedback might have disincentive faithful reasoning (Sharma, 2023). Models that are instruction-tuned or trained on the chain of thought (e.g., Flan-T5) during the pre-training phase have a better indirect effect across different reasoning tasks, suggesting that fine-tuning on CoT can make the model more faithful. Similar to Turpin et al. (2023), we observe inverse scaling for certain tasks. In our case, the indirect effect worsens with increasingly capable models, indicating that sheer scale might not guarantee faithful reasoning. Interestingly, we also observe that none of the models has high indirect or direct effects on the causal understanding task. One intuitive reason is that the causal understanding task is challenging, and the model's (<10B) performance is nearly random; hence, the effects are not strong. Overall, we observe that LLMs are inconsistent in faithfully performing reasoning over the CoT.

3 FRODO

In this section, we introduce FRODO, a framework that tailors small-sized LMs (<10B parameters) to be strong rationalizers and perform reasoning faithfully over the rationales. FRODO aims to improve the synergy between the reasoning chain and the final answer. We first describe how we obtain silver reasoning chains from LLMs (§3.1). Then, we introduce our inference module that trains a model to generate rationales (§3.1) followed by the reasoner module and its training objectives (§3.2).

3.1 Inference Module

In this work, we assume no gold rationales to train our model. Hence, similar to recent works (Liu et al., 2022, 2023; Wang et al., 2023; Ramnath et al., 2024), we automatically obtain the silver rationale from LLM (GPT-3) using in-context learning. A common approach is fine-tuning a smaller text-to-text model on the silver rationales generated

	Models	StrategyQA			GSM8k			Causal		
		CIE	CDE	p-value	CIE	CDE	p-value	CIE	CDE	p-value
AR	**LLaMA-2-7B**	24.5	25	<0.001	27.5	8.5	<0.001	2.3	1.1	<0.005
	Mistral-7B	21.2	17.9	<0.001	25.1	3.8	<0.001	2.3	0.6	<0.009
In-context	**LLaMA-2-7B**	24.9	10	<0.005	45.6	0.9	<0.005	5.6	5.6	<0.009
MoE	**Mixtral-8-7B**	21	11	<0.001	47.4	2.9	<0.003	5.1	4.6	<0.001
RLHF	**LLaMA-2-7B-Chat**	8.4	30.5	<0.010	1.4	36.7	<0.010	-2.3	8.6	<0.016
	Stable Vicuna-13B	3.5	2.5	<0.001	45.1	2.4	<0.010	0.6	0.1	<0.010
	ChatGPT	2.6	13.6	<0.016	57.8	16.6	<0.010	4.6	10.8	<0.001
Instruct Tuned	**Mistral-Instruct-7B**	31.6	31.9	<0.001	35.5	4.7	<0.001	7.4	8	<0.005
RLHF + Instruct Tuned	**GPT-3.5-Instruct**	26.1	27.3	<0.005	**62.6**	14.7	<0.005	**8.5**	10.7	<0.005
Instruct-Tuned + CoT Tuned	**Flan-T5-11B**	**36.9**	35.7	<0.001	31.23	12.2	<0.001	7.4	13.1	<0.001
	Flan-Alpaca-11B	31.2	47.9	<0.001	25	7.9	<0.001	3.4	9.2	<0.001

Table 3: **Causal Effects** of CoT. The reported results are zero-shot performance. CIE: Controlled Indirect Effect, CDE: Controlled Direct Effect. The p-value represents the significance of the results

by LLMs with a standard language modeling loss. Recent studies have shown that fine-tuning models (<5B) on reasoning chains may struggle to *align* the reasoning chains with the provided reasoning question during inference (Yang et al., 2023; Fu et al., 2023). Additionally, learning to generate a reasoning chain means learning to decompose complex reasoning into smaller reasoning steps implicitly. However, Shridhar et al. (2023) showed that fine-tuning could lead to learning shortcuts and degrade performance. Recent studies have demonstrated that feedback-based methods can help the model align better with the human goal. Hence, we use Direct Preference Optimization (DPO) (Rafailov et al., 2023) for aligning LMs to learn to generate correct reasoning chains.

Preference Data. We prompt the LLM to generate correct reasoning chains (R_w) and incorrect reasoning chains (R_l) for each reasoning problem. In our experiments, we consider two kinds of reasoning chains as incorrect: *counterfactual chains* (alternative chains that can lead to different outcomes) and *irrelevant chains*. We assume that models that can understand and learn to prefer correct reasoning chains over counterfactual chains will become more robust and enhance generalization. Hence, we manually construct correct and incorrect intermediate reasoning steps and demonstrate the model with these annotated examples before a new instance is provided. In this way, we obtain a preference data $D \in \{X, R_w, R_l\}$ that contains reasoning problems (X) and pairs of reasoning steps that lead to correct (Y_w) or incorrect outcomes (Y_l).

Training. Given a reasoning problem $\{x \in X\}$ and instruction prompt $p \in$ {correct or counterfactual}, our goal is to train models that could generate reasoning steps (r_w or r_l). We propose to adopt Direct Preference Optimization (DPO) (Rafailov et al., 2023), an effective algorithm for aligning language models with implicit rewards. DPO assumes that we only have access to some pairwise preference data $x \rightarrow \{r_w > r_l\}$ for each problem $x \in X$. Hence, while training a model (π_θ) to generate correct reasoning steps, we consider counterfactual and irrelevant reasoning steps as less preferred. Training a DPO model includes two phases: (i) supervised fine-tuning (SFT) and (ii) RL fine-tuning phase.

SFT. We begin by fine-tuning a pre-trained LM with a maximum log-likelihood objective to obtain π_{sft}.

RL Phase. Contrary to traditional RL approaches, which initially train a reward model and subsequently derive a policy from it, DPO enables extracting policy through implicit reward learning. The preference data of human or artificial annotators is modelled by a learnable implicit reward model f_θ under Bradley-Terry theories (Bradley and Terry, 1952):

$$\pi_\theta(r_w > r_l | x) = \sigma(f_\theta(r_w, x) - f_\theta(r_l, x)) \quad (1)$$

where σ is the sigmoid function. To learn f_θ, DPO adopts a binary classification loss:

$$L_{\text{DPO}} = -\mathbb{E}_{\{x, r_w > r_l\}} \log \sigma(f_\theta(r_w, x) - f_\theta(r_l, x)) \quad (2)$$

The latent function f_θ is parameterized by the log-likelihood ratio between π_θ and π_{sft}:

$$f_\theta(x, r) = \beta \log \frac{\pi_\theta(r|x)}{\pi_{\text{sft}}(r|x)} \quad (3)$$

where β a linear coefficient for scaling f_θ. This parameterization is appealing as it aligns the training of an implicit reward model f_θ closely with training an LM policy π_θ.

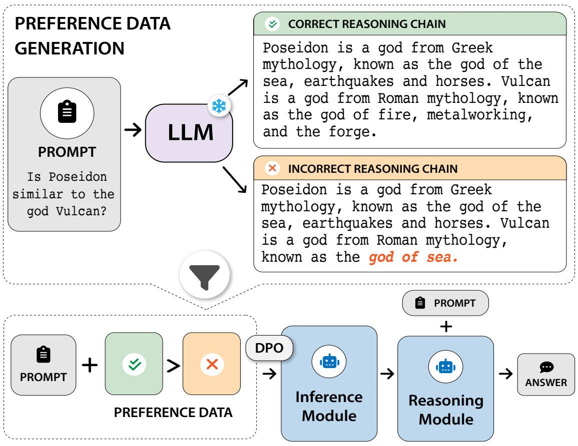

Figure 3: An overview of FRODO.

During inference, the reasoning module uses the generated reasoning steps by π_θ model for a given reasoning problem.

3.2 Reasoning Module

Given a reasoning question $x \in X$ and reasoning steps r_w (correct) and r_l (counterfactual)[4], our goal is to train a model (π_γ) that can generate a correct answer y_w. To encourage our reasoner module to reason faithfully over the reasoning steps, we train the model with a linear combination of two losses: an indirect effect loss and a supervised margin rank loss, $\mathcal{L} = \lambda_{\text{LM}} * \mathcal{L}_{\text{LM}} + \lambda_{\text{counter}} * \mathcal{L}_{\text{counter}} + \lambda_{\text{PREF}} * \mathcal{L}_{\text{PREF}}$, which we describe below.

Language Model Loss. We use the standard training objective to maximize the likelihood of the correct answer using cross-entropy loss, computed as:

$$\mathcal{L}_{\text{LM}} = -\log \text{P}(y_w | x, r_w) \tag{4}$$

Counterfactual Loss. To encourage the model to reason robustly and faithfully towards the reasoning steps, we propose training the model to learn how different reasoning chains (correct or counterfactual) can lead to different outcomes. Hence, inspired by the causal mediation theory (Pearl, 2001), we use the following loss:

$$\mathcal{L}_{\text{counter}} = -\log \text{P}(y_l | x, r_l) \tag{5}$$

Similar to (Wang et al., 2023; Roese, 1997), we posit that adding a counterfactual objective can help the model to avoid learning reasoning shortcut between a question and the gold answer since now the model is tasked to answer differently for the same question.

[4]Please note that in the reasoner module, we only consider counterfactual reasoning steps as negative samples.

Margin-Ranking Loss. It has been shown (Khosla et al., 2020) that contrastive loss and ranking loss help to improve model robustness and generalization against input variation. Hence, we propose to use the margin ranking loss that aims to maximize the margin between positive examples (i.e., statements containing questions, correct reasoning steps and correct answers) and negative examples (i.e., statements containing questions, counterfactual reasoning steps and correct answers).

$$\mathcal{L}_{\text{PREF}} = max(0, t * \text{IE} + m) \tag{6}$$

where t is the label (indicating which sample in the pair is better)=1, m is the margin=1.0, and the indirect effect IE = $h(x, r_w, y_w) - h(x, r_l, y_w)$ where h is the logits.

4 Experiments

Datasets. We conduct the causal mediation analysis on three datasets: STRATEGYQA (Geva et al., 2021), GSM8K (Cobbe et al., 2021), and Causal Understanding (Suzgun et al., 2023). STRATEGYQA and Causal Understanding are binary classification tasks. GSM8K dataset contains 8.5K high-quality linguistically diverse grade school math word problems. We evaluate FRODO on four datasets: STRATEGYQA, QUAREL(Tafjord et al., 2018), OPENBOOKQA (Mihaylov et al., 2018), and QASC (Khot et al., 2020). We report more details about each dataset in App. A.2. For all the datasets, we do not use human-written rationales. We used rationales generated by prior work (Ramnath et al., 2024) using GPT-3 (TEXT-DAVINCI-003) for silver rationales for supervision. For counterfactual rationales, we use chain-of-thought prompts on these datasets (refer App. A.5) and sample 2 rationale generations for each training set instance with a temperature of 0.5.

Evaluation Metrics. To evaluate the causal effects, we report the average indirect and direct effects of the LLMs. We use the following formula to calculate the scores: IE = Avg. Acc$(Y|X_0, R_0)$ − Avg. Acc$(Y|X_0, R_1)$, and DE = Avg. Acc$(Y|X_0, R_0)$ − Avg. Acc$(Y|X_1, R_0)$ where X_0 and R_0 original reasoning problem and reasoning chains. We measure two different kinds of causal effects: **natural** and **controlled** for different types of LLMs. *Natural Indirect Effect*: for models that have emergent capabilities (>100B parameters) of generating plausible reasoning chains, we measure the causal effect of X on Y that uses

R generated by the same model. *Controlled Indirect Effect*: for models with <20B parameters, we evaluate the causal effect by providing reasoning chains generated by GPT-4. Further, to measure the *robustness* of models, we use controlled indirect effect. To evaluate the *faithfulness* of the rationales generated by the small-sized models, we use LAS (Hase et al., 2020) to measure how well the rationales help a simulator to predict a student's predictions a', namely Acc(qr –> a') - Acc(q –> a'). Similar to Wang et al. (2023), we implement each simulator with a fine-tuned T5-large model respectively.

Implementation Details. We use GPT-4 to generate intervened reasoning problems X_1 and reasoning chains (R_0 or R_1) to perform the causal mediation analysis. We report the prompts used in App. A.5 and hyperparameters in App. A.2.

Baselines. We perform the causal analysis on a series of language models that are diverse in terms of scale, training, and data: LLaMa-2 (Touvron et al., 2023), Mistral (Jiang et al., 2023), ChatGPT (Brown et al., 2020a), GPT-4 (OpenAI, 2023), Flan-T5 (Chung et al., 2022), Flan-Alpaca (Chung et al., 2022), Stable-Vicuna (Chiang et al., 2023). We compare FRODO with four strong baselines: (1) SFT + CoT: Finetuning a T5-large or T5-3B or LLaMa-2-7B with LoRA or Mistral-7B with LoRA on silver rationales, then train another model with LM objective to perform the reasoning, Rainier (Liu et al., 2022), (3) Crystal (Liu et al., 2023), (4) Mario(Ramnath et al., 2024) and (5) SCOTT(Wang et al., 2023). More details about all the baselines are reported in App. A.3.

5 FRODO Results

Comparing FRODO with baselines. We now empirically compare FRODO with three strong baseline models (see Table 4). We consider T5-large (770M) as the inference and reasoning modules. We have the following three observations. First, we present the performance of GPT-3.5 on these tasks. We observe the performance on StrategyQA is much lower than on other tasks, indicating the rationales generated for this task can be unfaithful. Hence, similar to (Ramnath et al., 2024), for training FRODO, we use only the instances where the answer predicted by GPT-3.5 is correct. Second, for all four datasets, we observe that FRODO outperforms the strong self-rationalization baselines. FRODO, on average, improves the performance by +4.1 and +3 accuracy points compared to the SFT + CoT and MARIO (the strongest baseline), respectively, across all four tasks. Since SFT + CoT and MARIO use the same knowledge from GPT-3.5, our results suggest that both our inference and reasoning modules bring substantial performance gains to the model. Third, it is worth noting that increasing (770M to 3B) the model size does not hamper the performance of FRODO. Fourth, we also report the performance of the LLaMa-2-7B and Mistal-7B models. We show that FRODO further improves the performance of model size 7B.

Models	StrategyQA	QuaRel	OBQA	QASC
GPT-3.5$^{\diamond}$	69.7	83.4	84.5	80.3
SFT	57.6	74.6	65.0	58.6
SFT + CoT	63.6	77.7	65.5	59.4
Rainier	–	–	69.7	54.9
Crystal	–	–	64.2	56.8
MARIO	65.1	79.9	66.1	60.1
FRODO	**68.4***	**83.4***	70.2^{+}	**64.2***
-DPO	66.2	82.2	68.1	62.4
-CL	65.2	82.1	66.4	60.1
-MRL	65.5	81.3	66.2	62.1
SFT	63.1	81.29	72.0	67.8
SFT + CoT	65.1	84.2	73.3	72.0
SCOTT	61.5	–	–	65.0
Crystal	–	–	78.3	74.3
FRODO	**82.1***	**93.5***	**80.1***	**75.9***
LlaMa-2-7B	67.2	56.8	47.5	49.6
SFT + CoT	79.4	68.4	62.8	54.6
FRODO	81.5^{+}	73.5^{+}	71.4^{+}	63.4^{+}
Mistral-7B	58.2	56.8	82.1	65.2
SFT + CoT	78.2	70.8	83.5	70.1
FRODO	81.9^{+}	78.2^{+}	84.9^{+}	72.3^{+}

Table 4: Performance of small-sized LMs (770M-7B) on four different reasoning tasks. The base models are T5-large (770M), T5-3B (3B), LLaMa-2-7B and Mistral-7B. We report accuracy (%).$^{\diamond}$: few-shot performance, *: p-value<0.01, $^{+}$: p-value<0.05

Ablation. To obtain a better insight into the contribution of each component of FRODO, we perform an ablation study (see Table. 4). First, when we do not use the DPO to train our inference module, we see a consistent drop (-1.9%) in performance across the four tasks, indicating the importance of incorporating implicit feedback provided by the DPO in the model's training. Further, we observe a considerable drop in performance when we do not use counterfactual (-3.1%) and margin ranking loss (-2.8%). This result highlights the model's ability to benefit from including counterfactual examples.

6 Quantative Analysis

Robustness. In Table 5, we report the controlled indirect effect that indicates how robustly models are able to change their answers when provided with

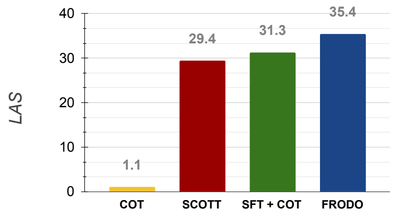

Figure 4: **Faithfulness** (LAS) of the compared methods on StrategyQA. The base Model is T5-3B.

controlled (generated by GPT-4) counterfactual reasoning chains. For STRATEGYQA, we observe that FRODO significantly improves the robustness performance for T5-3B (+7.7 pp.). Further, for the QuaRel task, we observe +2 pp. improvement over the SFT + CoT method. Qualitatively, we find that for the MCQA tasks, the gold rationales often contain the answer; hence, the SFT + CoT learns to copy those as answers. Further, we perform an ablation to understand which component contributes most to the model's robustness. We find that counterfactual loss brings the most gain in robustness.

Generalization. The idea is to test our model's capability to determine if it can improve out-of-distribution (OOD) generalization. Table 6 shows the OOD performance, where we compare our method with SFT+CoT. We trained the models on the OBQA and QASC datasets and evaluated them on the StrategyQA task. We conclude that FRODO significantly helps improve the model's generalizability to a dataset unseen during fine-tuning.

Models	StrategyQA	QuaRel
SFT	19.4	19.4
SFT + CoT	32.2	29.2
FRODO	**39.9**	**31.2**
-CL	34.6	28.7
-MRL	36.2	30.6

Table 5: **Robustness Performance** of LLMs on Reasoning over a Chain. We report CIE scores.

Faithfulness. Finally, we compare the faithfulness of reasoning chains generated by FRODO with SCOTT, COT and SFT+COT (see Fig.5). We observed that FRODO achieves a much higher LAS score than the other three baselines, suggesting that DPO training with implicit casual feedback helped the model.

7 Related Work

Measuring Faithfulness CoT. Jacovi and Goldberg (2020) argued that obtaining *faithful* explanations that accurately reflect a model's reasoning process is important to understand the reasons behind its answer.(Atanasova et al., 2023) proposed a new benchmark to test the faithfulness of natural language explanations. Turpin et al. (2023) proposed identifying examples of unfaithful CoT in adversarial settings, showing that CoT reasoning is not always faithful. To determine faithfulness, they provided bias features in the few-shot setting or made edits to the input. (Lanham et al., 2023) argued that LLM ignores mistakes when introduced into the CoT, which reveals that the LLM is unfaithful. Finally, (Parcalabescu and Frank, 2023) introduced CC-SHAP to measure input alignment with predictions for both post-hoc and CoT explanations. Unlike prior work, we employ causal mediation analysis to measure the model's faithful reasoning over the CoT, and to interpret its relationship with the answer.

Models	OBQA → SQA	QASC → SQA
T5-3B + CoT	67.6	53.2
FRODO	69.4	56.2

Table 6: **Generalization Performance** (accuracy) of methods, trained on a source dataset and directly predicting on a target dataset (denoted as source → target).

Self-Rationalization & CoT Distillation. Initial work on self-rationalization approaches focused on collected gold human rationales and training a model to learn to generate such rationales (Wiegreffe et al., 2021; Paul and Frank, 2021; Camburu et al., 2018). With the advent of LLMs, recently many works have distilled CoT from LLMs and endowed small LMs with step-by-step reasoning capabilities (Fu et al., 2023; LI et al., 2022; Shridhar et al., 2023; Li et al., 2023). We distill CoT similar to prior work, but further improve the correctness of the CoT using implicit feedback.

8 Conclusion

In this work, we perform a causal mediation analysis to study the indirect effect of CoT on the final output of twelve LLMs. Our experiments show large variations across tasks and models in how strongly reasoning traces causally affect the model's prediction. LLMs generally do not reliably use their intermediate reasoning steps when generating an answer. We introduce FRODO that tailors small-sized LMs to generate correct reasoning chains and faithfully reason over them to arrive at the correct answer. Experiments show that our method outperforms strong baselines on four reasoning tasks, including out-of-distribution settings.

9 Limitations

Compared to training a standard CoT distillation process, our method requires (i) additional counterfactual data generated by LLMs, which can be expensive, and (ii) training time increases as training Direct Preference Optimization is a two-step process. To manage the complexity of our already large-scale experiments involving (a) four different reasoning tasks, and (b) hyperparameter search grids, we ran experiments with 3 random seeds. Additionally, FRODO is dependent on rationales generated by LLMs. Extra care should be taken when applying our model in production environments, especially when making critical decisions or exposing its generated contents directly to human end users.

References

OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Benjamin Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Sim'on Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Raphael Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Lukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Ryan Kiros, Matthew Knight, Daniel Kokotajlo, Lukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Adeola Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel P. Mossing, Tong Mu, Mira Murati, Oleg Murk, David M'ely, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Ouyang Long, Cullen O'Keefe, Jakub W. Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alexandre Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Pondé de Oliveira Pinto, Michael Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya

Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario D. Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin D. Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas A. Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cer'on Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report.

Afra Feyza Akyurek, Ekin Akyürek, Aman Madaan, A. Kalyan, Peter Clark, D. Wijaya, and Niket Tandon. 2023. Rl4f: Generating natural language feedback with reinforcement learning for repairing model outputs. In Annual Meeting of the Association for Computational Linguistics.

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. 2023. Faithfulness tests for natural language explanations. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 283–294, Toronto, Canada. Association for Computational Linguistics.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs the method of paired comparisons. Biometrika, 39:324–345.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. Language models are few-shot learners. In Advances in Neural Information Processing Systems, volume 33, pages 1877–1901. Curran Associates, Inc.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020b. Language models are few-shot learners. ArXiv, abs/2005.14165.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, Advances in Neural Information Processing Systems 31, pages 9539–9549. Curran Associates, Inc.

Ziwei Chen, Linmei Hu, Weixin Li, Yingxia Shao, and Liqiang Nie. 2023. Causal intervention and counterfactual reasoning for multi-modal fake news detection. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 627–638, Toronto, Canada. Association for Computational Linguistics.

Yew Ken Chia, Pengfei Hong, Lidong Bing, and Soujanya Poria. 2023. Instructeval: Towards holistic evaluation of instruction-tuned large language models. arXiv preprint arXiv:2306.04757.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. Preprint, arXiv:2210.11416.

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. arXiv preprint arXiv:2110.14168.

Patrick Fernandes, Aman Madaan, Emmy Liu, António Farinhas, Pedro Henrique Martins, Amanda Bertsch, José GC de Souza, Shuyan Zhou, Tongshuang Wu, Graham Neubig, et al. 2023. Bridging the gap: A survey on integrating (human) feedback for natural language generation. arXiv preprint arXiv:2305.00955.

Matthew Finlayson, Aaron Mueller, Stuart M. Shieber, Sebastian Gehrmann, Tal Linzen, and Yonatan Belinkov. 2021. Causal analysis of syntactic agreement mechanisms in neural language models. ArXiv, abs/2106.06087.

Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. 2023. Specializing smaller language models towards multi-step reasoning. In Proceedings of the 40th International Conference on Machine Learning, volume 202 of Proceedings of Machine Learning Research, pages 10421–10430. PMLR.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. Transactions of the Association for Computational Linguistics, 9:346–361.

Peter Hase, Shiyue Zhang, Harry Xie, and Mohit Bansal. 2020. Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language? In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 4351–4367, Online. Association for Computational Linguistics.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4198–4205, Online. Association for Computational Linguistics.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. ArXiv, abs/2310.06825.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. Advances in neural information processing systems, 33:18661–18673.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. Qasc: A dataset for question answering via sentence composition. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 8082–8090.

Geunwoo Kim, Pierre Baldi, and Stephen Marcus McAleer. 2023. Language models can solve computer tasks. ArXiv, abs/2303.17491.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In Advances in Neural Information Processing Systems, volume 35, pages 22199–22213.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson E. Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, John Kernion, Kamil.e Lukovsiut.e, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Samuel McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, T. J. Henighan, Timothy D. Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Janina Brauner, Sam Bowman, and Ethan Perez. 2023. Measuring faithfulness in chain-of-thought reasoning. ArXiv, abs/2307.13702.

Hector J Levesque. 1986. Knowledge representation and reasoning. Annual review of computer science, 1(1):255–287.

Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. 2023. Symbolic chain-of-thought distillation: Small models can also "think" step-by-step. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2665–2679, Toronto, Canada. Association for Computational Linguistics.

SHIYANG LI, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jingu Qian, Baolin Peng, Yi Mao, Wenhu Chen, and Xifeng Yan. 2022. Explanations from large language models make small reasoners better. ArXiv, abs/2210.06726.

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, pages 8938–8958, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiacheng Liu, Ramakanth Pasunuru, Hannaneh Hajishirzi, Yejin Choi, and Asli Celikyilmaz. 2023. Crystal: Introspective reasoners reinforced with self-feedback. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 11557–11572, Singapore. Association for Computational Linguistics.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. arXiv preprint arXiv: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. Preprint, arXiv:2303.17651.

Ninareh Mehrabi, Palash Goyal, Christophe Dupuy, Qian Hu, Shalini Ghosh, Richard S. Zemel, Kai-Wei Chang, A. G. Galstyan, and Rahul Gupta. 2023. Flirt: Feedback loop in-context red teaming. ArXiv, abs/2308.04265.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. Preprint, arXiv:2303.08774.

Liangming Pan, Michael Stephen Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2023. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies. ArXiv, abs/2308.03188.

Letitia Parcalabescu and Anette Frank. 2023. On measuring faithfulness of natural language explanations. arXiv preprint arXiv:2311.07466.

Debjit Paul and Anette Frank. 2021. COINS: Dynamically generating COntextualized inference rules for narrative story completion. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 5086–5099, Online. Association for Computational Linguistics.

Debjit Paul, Mete Ismayilzada, Maxime Peyrard, Beatriz Borges, Antoine Bosselut, Robert West, and Boi Faltings. 2024. Refiner: Reasoning feedback on intermediate representations. In EACL, volume abs/2304.01904.

Judea Pearl. 1998. Graphs, causality, and structural equation models. Sociological Methods & Research, 27:226 – 284.

Judea Pearl. 2001. Direct and indirect effects. Probabilistic and Causal Inference.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Lidén, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023a. Check your facts and try again: Improving large language models with external knowledge and automated feedback. ArXiv, abs/2302.12813.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023b. Instruction tuning with gpt-4. arXiv preprint arXiv:2304.03277.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In Thirty-seventh Conference on Neural Information Processing Systems.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research, 21(1):5485–5551.

Sahana Ramnath, Brihi Joshi, Skyler Hallinan, Ximing Lu, Liunian Harold Li, Aaron Chan, Jack Hessel, Yejin Choi, and Xiang Ren. 2024. Tailoring self-rationalizers with multi-reward distillation. In The Twelfth International Conference on Learning Representations.

Neal J. Roese. 1997. Counterfactual thinking. Psychological bulletin, 121 1:133–48.

Gauri Sharma. 2023. Discovering safety issues in text-to-image models: Insights from adversarial nibbler challenge. In Proceedings of the ART of Safety: Workshop on Adversarial testing and Red-Teaming for generative AI, pages 43–48, Bali, Indonesia. Association for Computational Linguistics.

Noah Shinn, Federico Cassano, Beck Labash, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. 2023. Distilling reasoning capabilities into smaller language models. In Findings of the Association for Computational Linguistics: ACL 2023, pages 7059–7073, Toronto, Canada. Association for Computational Linguistics.

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, pages 7035–7052, Singapore. Association for Computational Linguistics.

Zhiqing Sun, Xuezhi Wang, Yi Tay, Yiming Yang, and Denny Zhou. 2023. Recitation-augmented language models. In The Eleventh International Conference on Learning Representations.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In Findings of the Association for Computational Linguistics: ACL 2023, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics.

Oyvind Tafjord, Peter Clark, Matt Gardner, Wen tau Yih, and Ashish Sabharwal. 2018. Quarel: A dataset and models for answering questions about qualitative relationships. In AAAI Conference on Artificial Intelligence.

Graham Tierney and Alexander Volfovsky. 2021. Sensitivity analysis for causal mediation through text: an application to political polarization. In Proceedings of the First Workshop on Causal Inference and NLP, pages 61–73, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. 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, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. ArXiv, abs/2307.09288.

Miles Turpin, Julian Michael, Ethan Perez, and Sam Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. ArXiv, abs/2305.04388.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In Advances in Neural Information Processing Systems, volume 33, pages 12388–12401. Curran Associates, Inc.

Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. 2023. SCOTT: Self-consistent chain-of-thought distillation. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5546–5558, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. CoRR, abs/2201.11903.

Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. 2021. Measuring association between labels and free-text rationales. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 10266–10284, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.

Guang Yang, Yu Zhou, Xiang Chen, Xiangyu Zhang, Terry Yue Zhuo, and Taolue Chen. 2023. Chain-of-thought in neural code generation: From and for lightweight language models. ArXiv, abs/2312.05562.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In The Eleventh International Conference on Learning Representations.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2023. How language model hallucinations can snowball.

A Appendix

A.1 More Related Work

Measuring Faithfulness CoT. Jacovi and Goldberg (2020) argued that obtaining *faithful* explanations that accurately reflect a model's reasoning process is important to understand the reasons behind its answer. Recently, Atanasova et al. (2023) proposed a new benchmark to test the faithfulness of natural language explanations. Turpin et al. (2023) proposed identifying examples of unfaithful CoT in adversarially constructed settings, showing that CoT reasoning is not always faithful. To determine faithfulness, they provided bias features in the few-shot setting or made edits to the input. (Lanham et al., 2023) argued that LLM ignores mistakes when introduced into the CoT, which reveals that the LLM is unfaithful. Finally.(Parcalabescu and Frank, 2023) proposed CC-SHAP to measure how well a model's input contributions align when it produces a prediction and explanation and use it for post-hoc and CoT explanations. Unlike previous work in this paper, we propose measuring the model's faithful reasoning over the CoT. We propose to use causal mediation analysis to interpret the relationship between the CoT traces and the final answer.

Self-Rationalization & CoT Distillation. Initial work on self-rationalization approaches focused on collected gold human rationales and training a model to learn to generate such rationales (Camburu et al., 2018; Wiegreffe et al., 2021; Paul and Frank, 2021). With the advent of LLMs, recently many works have distilled chain-of-thought from LLMs and endowed small LMs with step-by-step reasoning capabilities (Fu et al., 2023; LI et al., 2022; Shridhar et al., 2023; Li et al., 2023). Our work involves distilling CoT from LMs to a smaller one, similar to a certain line of work. We differ in using implicit feedback to enhance the correctness of the distilled CoT.

Feedback to Improve Reasoning. Recently, several papers have proposed to improve or revise the LMs' generation using feedback (Fernandes et al., 2023; Pan et al., 2023). Broadly, existing methods can be categorized into two kinds: external and intrinsic feedback. In the realm of external feedback, a standard procedure is to train critic models and use them to facilitate and improve the original generation model (Peng et al., 2023a; Akyurek et al., 2023; Mehrabi et al., 2023; Paul et al., 2024). Among them, Paul et al. (2024) is related to our paper as it evaluates each reasoning step as feedback to produce more reasonable reasoning steps. In contrast to extrinsic feedback, which relies on external sources, there are works which show that internal knowledge of LLMs can be used to give feedback (Kim et al., 2023; Madaan et al., 2023; Shinn et al., 2023). However, Madaan et al. (2023) argued that self-feedback does not improve performance on reasoning tasks. Hence, in this work, we create preference data (counterfactual and factual reasoning steps) to train a specialized model to learn to generate correct reasoning steps with implicit feedback.

Casual Mediation Analysis in NLP. Causal mediation analysis is an important tool that is used to effectively attribute the causal effect of mediators on an outcome variable (Pearl, 2001). Vig et al. (2020) proposed to use this method to implicate specific neurons and attention heads in mediating gender bias in various pre-trained LMs. Later, this method was used for analyzing different models' behaviour for different downstream tasks such as Subjective-Verb agreement (Finlayson et al., 2021), Fake News Detection (Chen et al., 2023), arithmetic reasoning (Stolfo et al., 2023), political polarization (Tierney and Volfovsky, 2021). To the best of our knowledge, our study is the first attempt to use casual mediation analysis to analyze the faithfulness of LLMs in their reasoning capabilities. In this work, we followed Pearl (2001) to perform the mediation analysis. The mediation analysis allows us to measure the following: Direct effect: Contribution of X (input) to Y (output). Indirect effect: Contribution of R (reasoning chain) to Y (output). Hence, a high direct effect means the model's output (Y) is primarily influenced by the input (X), and a high indirect effect means the reasoning chain (R) has more effect on the output (Y).

A.2 Dataset and Implementation Details

All datasets have multi-choice questions "*yes*/*no*" for STRATEGYQA, "*a*/*b*" for QUAREL, "a/b/c/d" for OPENBOOKQA, "*a*/*b*/-/*h*" for QASC), and the task is to generate a rationale followed by the predicted answer. We use the original data splits (see Table.10).

Data Size	Test Data Size
GSM8K	300
Causal Understanding	175
StrategyQA	500

Table 7: Data Statistics: Causal Mediation Analysis

Hyperparameter	Value
Optimizer	Adam
Adam epsilon	$1e-8$
Adam initial learning-rate	$3e-5$
Learning-rate scheduler	linear with warmup
Warmup steps	1000
Gradient clipping	0.5
Train batch-size	4/8
Training Time	$\sim$ 4 hours on 1 GPU

Table 8: Training Details for small LMs

Hyperparameter	Value
Optimizer	RMSprop
Adam epsilon	$1e-8$
Train batch-size	4/8
beta	0.25
Training Time	$\sim$ 8 hours on 1 GPU
LoRA parameters	
task type	CAUSALLM
r	16
lora alpha	32
lora dropout	0.05

Table 9: Training Details for Direct Preference Optimization

A.3 Baselines

We evaluate a series of language models that are diverse in terms of scale, training, and data:

- **LLaMA** (Touvron et al., 2023), an open-source decoder-only model with various sizes (7B) model is pretrained using only a language modeling loss.

- **GPT-3.5** (Brown et al., 2020a) and **GPT-4** (OpenAI, 2023): two closed-source decoder-only models that were trained with instruction-tuning. For GPT-3.5, we use the `text-davinci-003` model with 175B parameters.

- **Stable-Vicuna**: open-source decoder-only model based on LLaMA. Stable-Vicuna is fine-tuned with RLHF.

- **Flan-T5-XXL** (Chung et al., 2022, 11B parameters) and **Flan-Alpaca** (Chia et al., 2023; Peng et al., 2023b; 3B), two open-source encoder-decoder models based on T5 (Raffel et al., 2020) and trained on instruction-following datasets.

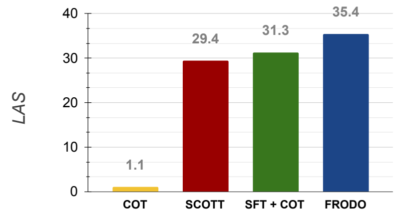

Figure 5: **Faithfulness** (LAS) of the compared methods on StrategyQA. The base Model is T5-3B.

A.4 Details about Preference Data

In our experiments, we consider two kinds of reasoning chains incorrect: counterfactual chains (alternative chains that can lead to different outcomes) and irrelevant chains (irrelevant facts about the correct answer). Additionally, we train DPO with a setting where for each instance, one correct and one incorrect chain are paired and given to the model. The model learns to prefer the correct chain. In Table 15, we report the size of the preference data used to train the DPO models.

A.5 Prompts

A.6 Causal Intervention

Dataset/Tools	Citation	Link	License
GSM8k	Cobbe et al. (2021)	https://github.com/openai/grade-school-math	MIT License
HuggingFace	Wolf et al. (2020)	https://github.com/huggingface/transformers	Apache License
OBQA	(Mihaylov et al., 2018)	https://huggingface.co/datasets/openbookqa	Apache License
StrategyQA	(Geva et al., 2021)	[1]	MIT License
Quarel	(Tafjord et al., 2018)	https://github.com/allenai/unifiedqa	MIT License
QASC	(Khot et al., 2020)	https://github.com/allenai/unifiedqa	MIT License

Table 10: More details about datasets and Tools [1] https://github.com/eladsegal/strategyqa/tree/main/data/strategyqa

{
"correct_answer": "False",
"counter_answer": " True",
"counter_question": "Would Finding Nemo be appropriate for a preschool class? ",
"counter_reasoning_chain": "Step 1: Finding Nemo is a Pixar animated film that is rated G, meaning it is suitable for all ages. Step 2: The film's themes and content are generally considered appropriate for young children, including preschoolers.",
"original gold facts": "Preschoolers are between 3 and 5 years old. The Dark Knight is rated PG-13. PG-13 is a rating that means parents are strongly cautioned that the content of a film may not be appropriate for children under 13",
"pred_answer_gpt4": "Step 1: The Dark Knight is a movie with intense action and violence.
Step 2: Preschool children are usually between the ages of 3 and 5 and are sensitive to violent content.
Step 3: Considering the violent content in The Dark Knight and the age group of preschool children, it would not be appropriate for a preschool class. Answer: False.",
"question": "Would The Dark Knight be appropriate for a preschool class?",
"GPT-4 prediction": "False"
}

Figure 6: Example from StrategyQA dataset. We can see one such instance where GPT-4 predicted the original answer given an incorrect reasoning chain.

Model	GSM8K
LLama-2 7B + SFT + COT	17.8
LLama-2 7B + SFT + FRODO	21.1
Mistral + SFT + COT	40.4
Mistral + SFT + FRODO	44.6

Table 11: Performance of FRODO on GSM8K (accuracy)

Model	Strategyqa → Causal Understanding
LLama-2 7B - SFT + COT	51.0
LLama-2 7B - SFT + FRODO	53.2

Table 12: Performance of FRODO on GSM8K (accuracy)

Dataset	Question	Option	Correct Answer	GPT-3 Generated CoT
StrategyQA	Can I build a house on an asteroid?	Yes or No	No	Building a house on an asteroid is impossible as of now due to the lack of technologies and resources needed. It would be extremely difficult to build a house that could withstand the extreme temperatures, radiation, and extreme gravitational pull.
OBQA	The circulatory system brings oxygen to the body from where?	(a) The brain (b) The feet (c) The stomach area (d) The chest	The chest	The circulatory system brings oxygen to the body from the lungs which is located in the chest area.
Quarel	The boys were racing their cars in the soapbox derby and found that the cars that − − − − moved faster.	(A) weighed more (B) weighed less	weighed less	When something is lighter, it is easier to move faster. Thus, the cars that weighed less moved faster.
QASC	What type of water formation is formed by clouds?	(A) pearls (B) streams (C) shells (D) diamonds (E) rain (F) beads (G) cooled (H) liquid	Beads	Rain is formed when water droplets in the clouds come together to form larger droplets that are too heavy to remain suspended in the cloud, and fall to the ground as precipitation.

Table 13: Examples from each reasoning task.

PROMPT: Counterfactual Prompts

```
You are a helpful assistant for commonsense reasoning.
We will provide you with a commonsense question, along with a correct
 answer and your task is to generate the counterfactual intermediate
steps. Here are two examples:

“Question : ” <Problem Statements> Let’s think step by step
<equation> Answer: <answer>

“Question: ” <Problem Statements> Let’s think step by step
<equation> Answer: <answer>

“Question: ” <Problem Statements> Let’s think step by step
```

Table 14: Prompts used for generating counterfactual intermediate reasoning steps.

Data type	StrategyQA	QuaRel	OBQA	QASC
Correct Reasoning Chain (R_w)	5492	8203	20138	19935
Counterfactual Reasoning Chain (R_l)	5492	8203	20138	19935
Irrelvant Reasoning Chain (R_l)	5492	8203	20138	19935

Table 15: Preference Data Statistics.

Tasks	Interventions
StrategyQA	Prompt GPT-4 to generate alternative questions such that the answer changes from original to counterfactual.
GSM8K	We automatically replace the operands with alternative operands.
Causal Understanding	Prompt GPT-4 to generate alternative questions such that the answer changes from original to counterfactual.

Table 16: Causal Interventions

Tasks	Generations
Question	Is the Illuminati card game still popular?
Gold Reasoning Chain	The original version of the game was released in 1982. A collectible card game version was released in 1995 but only had one set. The most recent edition of the base game was published in 2007.
SFT + CoT	The Illuminati card game was released in the 1980s.
DPO	The Illuminati card game was released in the 1980s. The Illuminati card game was discontinued in the 1990s.
correct answer	False

Table 17: Qualitative Examples of Outputs

{
    "New Answer": "108",
    "Alternative Question": "Janet\u2019s ducks lay 34 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder at the farmers' market daily for $4 per fresh duck egg. How much in dollars does she make every day at the farmers' market?",
    "Intervented_CoT": [
        "Janet sells 34 - 3 - 4 = 27 duck eggs a day.",
        "She makes 27 * 4 = $108 every day at the farmer\u2019s market."
    ],
    "Original answer": "18",
    "Original_question": "Janet\u2019s ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder at the farmers' market daily for $2 per fresh duck egg. How much in dollars does she make every day at the farmers' market?",
    "GPT-4_generated_CoT": [
        "Janet sells 16 - 3 - 4 = 9 duck eggs a day.",
        "She makes 9 * 2 = $18 every day at the farmer\u2019s market."
    ],
    "Gold reasoning_chain": "Step 1: Janet gets 16 eggs from her ducks every day. \nStep 2: She uses 3 for breakfast and 4 for baking muffins, so she uses 3 + 4 = <<3+4=7>>7 eggs per day.\nStep 3: Subtract the number of eggs she uses from the total number of eggs her ducks lay to find out how many she sells at the market: 16 - 7 = <<16-7=9>>9 eggs.\nStep 4: She sells her eggs for $2 each, so her daily income from egg sales is 9 * $2.",
    "GPT-4 predicted answer when given Original Question and incorrect chain": "18 dollars"
}

Figure 7: Example from GSM8K dataset. We can see one such instance where GPT-4 predicted the original answer given an incorrect reasoning chain.

{
"correct_answer": "True",
"counter_answer": " False",
"counter_question": "Could a wandering albatross fly from Bucharest to Sydney without a rest? ",
"counter_reasoning_chain": "Step 1: The wandering albatross is known for its ability to fly long distances without resting, often covering thousands of kilometers.\nStep 2: The distance from Bucharest, Romania to Sydney, Australia is approximately 15,000 kilometers.\nStep 3: The maximum recorded flight distance of a wandering albatross is about 15,000 kilometers, so theoretically, it could make the journey without resting.",
"counterfactual": "If the distance from Bucharest to New York City was more than 15,000 km, a wandering albatross would not be able to fly this distance without a rest.",
”golden facts": "Wandering albatross can travel at least 15,000 km (just under 10,000 miles) over the sea before returning to land. It's 4766 miles or 7670 km from Bucharest to New York City.",
"pred_answer_gpt4": "Step 1: The distance between Bucharest and New York City is approximately 7,500 kilometers (4,660 miles). Step 2: A wandering albatross can fly long distances, with a maximum range of up to 15,000 kilometers (9,320 miles) without needing to land. Step 3: Since the distance between Bucharest and New York City is less than the maximum range of a wandering albatross, it could fly from Bucharest to New York City without a rest.",
"question": "Could a wandering albatross fly from Bucharest to New York City without a rest?",
”GPT-4_y_01": "False”
”GPT-4_y_00": ”True"
}

Figure 8: Example from StrategyQA dataset. We can see one such instance where GPT-4 predicted the counter answer given an incorrect reasoning chain.