How Interpretable are Reasoning Explanations from Prompting Large Language Models?

Anonymous ACL submission

Abstract

Prompt Engineering has garnered significant attention for enhancing the performance of large language models across a multitude of tasks. Techniques such as the Chain-of-Thought not only bolster task performance but also delineate a clear trajectory of reasoning steps, offering a tangible form of explanation for the audience. Prior works on interpretability assess the reasoning chains yielded by Chain-of-Thought solely along a singular axis, namely faithfulness. We present a comprehensive and multifaceted evaluation of interpretability, examining not only faithfulness but also robustness and utility on 3 commonsense reasoning datasets. Likewise, our investigation is not confined to a single prompting technique; it expansively covers a multitude of prevalent prompting techniques employed in large language models, thereby ensuring a wide-ranging and exhaustive evaluation. In addition, we introduce a simple alignment technique, termed Self-Entailment-Alignment Chain-of-thought, that yields more than 70% improvement across all dimensions of interpretability. Our findings suggest that interpretability should be assessed from various dimensions instead of grounding our conclusions based on a singular metric.

1 Introduction

In recent trends, Large Language Models (LLM) have shown impressive performance across a diverse array of tasks, primarily through extensive scaling of model size (Brown et al., 2020). Techniques such as instruct-tuning (Wei et al., 2021) applied across diverse tasks have empowered LLMs to execute inference on previously unseen tasks.

One of the leading factors can be attributed to customizing the prompt to align with the specific targeted task. Given the considerable potential this holds for enhancing task performance, substantial research efforts have been channeled toward innovating newer ways of prompting LLMs to utilize their pre-training knowledge in a more effective manner.

Chain-of-Thought (CoT) (Wei et al., 2022) has gathered much attention due to its simple setup which allows the LLM to generate not only the task output but also the steps undertaken. In addition to its efficacy in enhancing the model's performance, this prompting method concurrently touches on one of the important aspects of utilizing these models for decision-making: interpretability.

The assumption is that the reasoning chain preceding the answer illustrates the model's thought process, enabling the audience to understand how the answer is derived. However, such claims though seemingly plausible should be taken lightly as they may not be faithful to the model's reasoning process (Jacovi and Goldberg, 2020). In this context, *plausibility* refers to the extent to which an explanation resonates with and is deemed acceptable by a human audience. *Faithfulness*, on the other hand, is characterized by the extent to which the explanation accurately reflects the model's decision-making process.

There has been a large number of works that seek to introduce modifications to CoT, including Self-Consistency (Wang et al., 2022b) and Least-to-Most (Zhou et al., 2022). We introduce a simple extension to the list of CoT variants, but purely with a focus on enhancing interpretability in the reasoning chain. The approach coined Self-Entailment-Alignment CoT (SEA-CoT) similarly utilizes a form of consistency between the set of possible outcomes, with an additional touch of alignment towards desirable explainability qualities.

Moreover, we conduct an extensive investigation into the reasoning explanations by evaluating under three pivotal axes of interpretability: faithfulness, robustness, and utility on 3 commonsense reasoning datasets. These assessments are implemented across multiple prompting techniques including CoT and various adaptations of it.

2 Motivation

Efforts aimed to enhance faithfulness in NLP take various forms. Extractive rationalizing model (Lei et al., 2016), designed to be faithful, generally comprises two separate components: explainer and predictor. This design paradigm conditions the predictor exclusively on text spans extracted by the explainer, positing that the resultant output, $\hat{y}$ is faithfully aligned with the extracted text, $\hat{e}$. However, prior studies (Wiegreffe et al., 2020) cautions against such beliefs, identifying limitations in adopting the explain-then-predict approach. The author mentions that such an approach restricts the focus of the predictor toward the target identified by the explainer, thereby raising questions about what is being explained. Conversely, Jacovi et al. (Jacovi and Goldberg, 2021) highlight concerns relating to the lack of meaningful insights from multiple text spans.

In accordance, we note that besides the limitation of narrowing the predictors' context, employing separate models could compromise producing a faithful explanation. As a start, we conduct a preliminary study, wherein we compare the faithfulness and utility of a single LLM that jointly predicts both $\hat{y}$ and $\hat{e}$, against another modular approach that involves two distinct LLMs, each tasked with predicting one of the two variables. We adopt the PINTO framework (Wang et al., 2022a), which uses an LLM, r_θ as the explainer while employing a smaller predictor, f_ϕ to generate the task label, $\hat{y} = f_\phi(x \oplus \hat{e})$ over the produced explanation, $\hat{e} = r_\theta(x)$ concatenated with the context. More importantly, PINTO addresses the label-specific issue by generating an explanation for each given option in a multiple-choice setup.

We are interested to see if generating both rationale and answer with a single model, yields better $\hat{e}$. In this setup, we train f_ϕ to generate both $\hat{e}$ and $\hat{y}$ jointly. We measure faithfulness by computing the drop in performance when swapping $\hat{e}_i$ with another instance within the same batch, $\hat{e_{j \neq i}}$ before deriving $\hat{y}|x; \hat{e}$. We use Leakage-Adjusted Simulatability (LAS) (Hase et al., 2020) to measure the utility of the rationale, a higher score would indicate that $\hat{e}$ is more useful towards learning $\hat{y}$.

We conduct experiments on two commonsense reasoning datasets: Commonsense QA (CSQA) (Talmor et al., 2018) and OpenBookQA (OBQA) (Mihaylov et al., 2018). Figure 1 shows that the joint approach scores higher on both accounts of faithfulness and utility. We hypothesize that a single model is in better control of aligning its explanation to the resultant outcome. Contrarily, a model relying on explanations synthesized by an external model may instead exhibit a diminished correlation between the interdependent variables, explaining the marginal difference in performance despite given an unrelated stimulus.

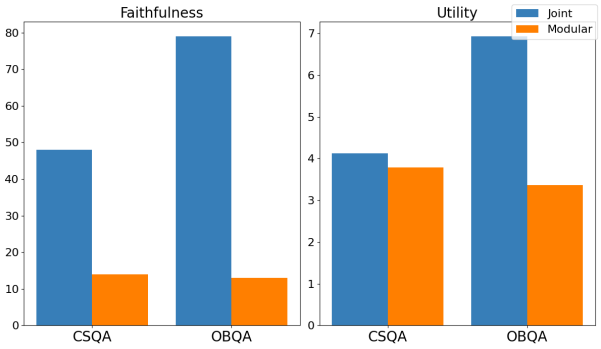

Figure 1: Faithfulness and Utility scores for joint and modular approach on two reasoning datasets: CSQA and OBQA.

Notably, this observation resonates well with the recognized capability of recent LLMs to autonomously generate text serving diverse objectives. In particular, LLMs pre-trained on a large amount of text can elucidate their reasoning processes, assisted with the appropriate prompting format. However, despite their apparent plausibility to human users, the quality of these explanations remains to be comprehensively validated.

3 Prompt Techniques

In this section, we systematically review various ways a LLM can be prompted. These methods primarily differ in how the language model is queried to derive the final answer, while the proposed approach focuses on deriving the final explanation. A high-level overview is shown in Figure 2.

- **CoT**: Chain-of-thought prompting has shown promising results in encouraging an LLM to better answer the task by reasoning aloud the steps before arriving at the final answer. (Kojima et al., 2022) has shown that it is possible in the zero-shot setting simply by appending *"Let's think step by step"* at the end of the instruction.

- **Self-Consistent CoT (SC-CoT)**: Following on, other works like Self-Consistency (Wang et al., 2022b) address the suboptimality of

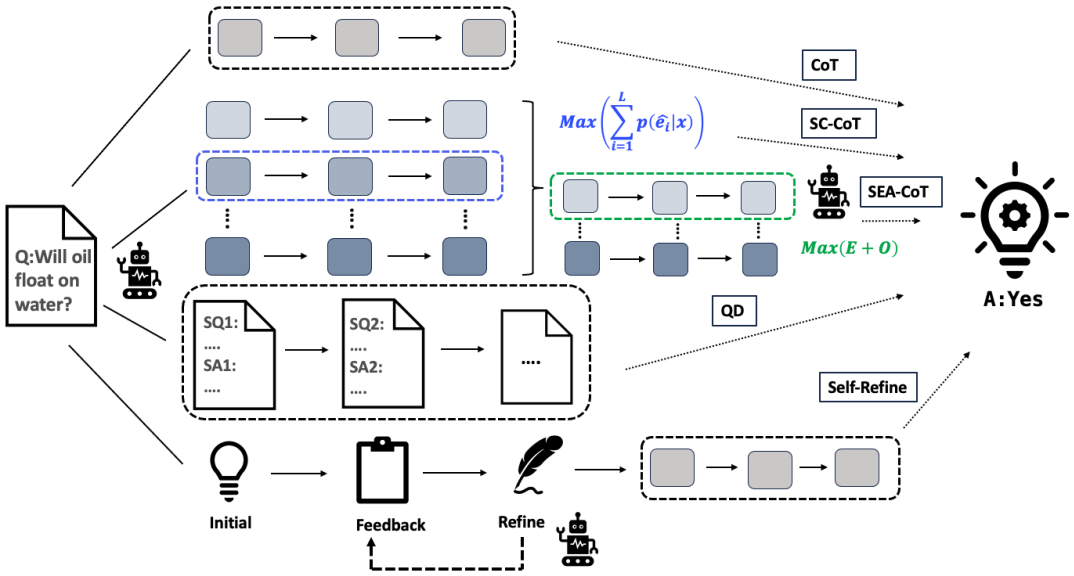

Figure 2: Overview of different prompting techniques to derive the reasoning chain, to serve as the explanation (boxed with dashed line). [Top to Bottom]: Cot, SC-CoT, SEA-CoT, QD, SR. SC-CoT and SEC-CoT differ in the explanation selection stage, where the former selects based on maximum cumulative probability (blue) and the latter (green) on two objectives: entailment, E, and overlap,O with an additional forward pass. Each robot figure denotes a forward pass from the LLM, SR stops when encounters a stopping criteria or exceeds the max number of passes.

greedy decoding in CoT by generating multiple paths and choosing the final answer, $\hat{y}^*$ via majority voting. SC-CoT has shown improvements across multiple arithmetic and commonsense reasoning benchmarks. Since multiple explanations may lead to the majority answer, we pick the one with the highest cumulative probability.

- **Question decomposition (QD)**: (Zhou et al., 2022) demonstrates that decomposing a complex problem into more manageable subproblems significantly facilitates the problem-solving capability of the model. The model answers each sub-problem and pieces together the answers to arrive at the final conclusion for the principal problem. We treat the sub-question and answers as the target explanation and assess their interpretability properties.

- **Self-Refine (SR)**: SR (Madaan et al., 2023) is a type of iterative process of prompting the LLM with a set of instructions. The main idea is to instruct the LLM to continuously provide feedback for its' own output and refine using the feedback, the process stops when the feedback deems the output as sufficient in solving the task at hand. The whole iterative process is achieved by self-prompting the same language model. There exist other forms of acquiring feedback, such as querying a trained feedback model or using external factual knowledge (Pan et al., 2023). We choose the approach of querying the same LLM as we are focused on the explainability of generated outputs from a sole LLM.

- **Self-Entailment-Alignment CoT (SEA-CoT)**: SEA-CoT is an adaptation from SC-CoT, supplemented with an additional ranking step to prioritize the top desirable reasoning explanation. Instead of selecting the most probable explanation, the reasoning is chosen based on the maximization of two objectives: entailment and the overlapping between Q&A, $(x \oplus \hat{y})$ and reasoning $\hat{e}$. We posit that a credible explanation should intrinsically align with the given context it aims to elucidate; in this scenario, it encompasses both the question being addressed and the predicted label. Maximizing the overlap between two sets of tokens can be seen as a measure of generating factual explanations, which concurrently aligns with the notion of faithfulness. Inspired by works that employ the LLM itself to do self-correction, we do the same by asking the LLM to rate the entailment level between its own generated reasoning and the joint context,

$x \oplus \hat{y}$. The LLM chooses between two options, entailment and contradiction. We then combine the probability of entailment together with the Intersection over Union (IoU) score, IoU($x \oplus \hat{y}, \hat{e}$). This approach is applicable only in the event where $|\hat{y}^*| = K > 1$, else we fall back to SC-CoT, though we note that this can be avoided by trivially setting the number of sequences, to be higher than the number of possible options, $N > |y|$.

4 Interpretability Qualities

Interpretability is a multifaceted characteristic and has multiple desirable traits with respect to the goal of the explanation (Yeo et al., 2023). One such goal can be instilling trust in the decision given by a model or another relating to understanding more about how the decision is derived. These attributes are not mutually exclusive and exhibit intersecting prerequisites. For instance, comprehending the decision-making process may foster trust in the system. This trust, in turn, can lead the user to perceive the decisions as being made on fair grounds, further reinforcing the user's confidence in the system. In our work, we focus on three aspects of interpretability: faithfulness, robustness, and utility.

4.1 Faithfulness, Robustness and Utility

The concept of faithfulness seeks to gauge the extent to which the explanation aligns with the underlying decision-making process. (Lanham et al., 2023) conducted a series of faithfulness tests, prompting the LLM with CoT. The objective of these tests is to introduce specific perturbations in the post-hoc explanations, and subsequently, scrutinize any resultant change in task outputs. This evaluation encompasses operations such as truncation of the reasoning chain at diverse lengths, paraphrasing, and intentional error introduction. In a parallel effort, we too employed paraphrasing and error introduction methodologies to assess the faithfulness of our model. We additionally employ counterfactual reasoning in our faithfulness assessment. Robustness, on the other hand, seeks to measure how resilient or consistent a given explanation is under various circumstances. For instance, employing adversarial attacks on an explanation, as delineated by (Chen et al., 2022), could serve as a mechanism to ascertain whether the model's decision is susceptible to diversion or distraction induced by these attacks.

Both faithfulness and robustness contribute to fostering trust and confidence in decisions made by an LLM. A faithful explanation facilitates a high level of trust among affected stakeholders and provides a means of identifying undesirable biases in the decision-making process. On the other hand, a robust explanation bolsters user confidence by assuring that the model is acting in the intended manner. Yet another under-studied axis of interpretability is the usefulness of the explanation. A useful explanation can facilitate knowledge transfer, resulting in benefits such as distillation in smaller models, debugging, or inspiration for self-improvement when presented to a human audience. Utility can be viewed as analogous to plausibility from a human's perspective since individuals naturally consider an explanation as plausible if it is useful in aiding them in understanding the decision made. We illustrate an overview of the perturbations in Figure 3.

4.2 Paraphrase

This assessment allows us to explore the intersection of robustness and faithfulness within the model's behavior. A faithful explanation, in alignment with the answer, should consistently mirror a similar decision-making process, leading to identical conclusions when presented with similar instances. Simultaneously, an explanation can be said to be robust if it enables slight re-wording of key inputs and still holds the same message when utilized by a model to make decisions. We utilize OpenAI's GPT3.5 to rephrase the target reasoning explanation, $\hat{e}$.

4.3 Adding mistakes

In contrast to ensuring answer consistency among similar reasoning, we conducted another test by adding mistakes to the reasoning before requesting the answer from the target LLM. One would expect the model to change its decision given an erroneous reasoning chain. We note that the focus in this context lies with the alteration in prediction rather than actual task performance. Since an incorrect reasoning may potentially transition to a correct one upon the introduction of an error, albeit such occurrences are exceedingly rare. Similarly, GPT3.5 is used to add non-factual errors to the provided reasoning.

4.4 Simulatability

Since it is costly to employ humans to assess if a reasoning chain is said to be useful, we employ

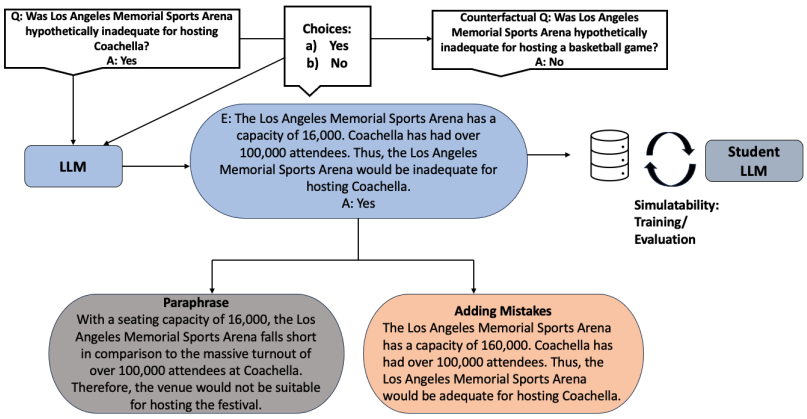

Figure 3: Interpretability test for faithfulness, robustness and utility. Perturbations to reasoning explanation: paraphrase and adding mistakes. Perturbations to context: counterfactual reasoning.

simulatability as a proxy for utility. We measure simulatability using LAS in Section 2 as it has been shown to be highly correlated with human judgment. A 220M T5-base (Raffel et al., 2020) is selected as the student model. The generated reasoning, $\hat{e}$ is appended to the input context x, which is then used as the final context for predicting the task label, $\hat{y} = f_s(\hat{e} \oplus x)$, where f_s refers to the student model. The student model undergoes fine-tuning with the aid of these samples, followed by an evaluation of its performance. A key aspect of LAS lies with the notion of subtracting a baseline, $f_s(x)$ from $f_s(\hat{e} \oplus x)$. This is used to simulate the additional benefits gained by using $\hat{e}$ in the training process to infer y.

4.5 Counterfactual reasoning

An alternative method to ascertain faithfulness follows by evaluating whether an explanation would change when the original question is modified in a different direction, particularly when directed towards a counterfactual scenario. (Atanasova et al., 2023) shows that an instance of unfaithfulness can be detected if the counterfactual explanation, e' does not acknowledge the modifications, c in the counterfactual instance $x'_i : y'$, yet still successfully predicting the counterfactual label, $y' \neq y$. Such an occurrence would mean that the counterfactual explanation is not faithfully aligned with the answer it supports. The distinction from Section 4.3 is that besides detecting signs of unfaithfulness, it also embodies a directed approach that assesses a model's capacity to contemplate alternative scenarios. Conversely, introducing mistakes serves as an undirected measure aimed at gauging the decline in confidence regarding the consistency of the model's output, without specifically targeting the assessment of the model's knowledge base. Specifically, we deemed an instance as unfaithful under the following conditions:

1. $x'_i = \{x_{i,1}, x_{i,2}...c, ...x_{i,L}\} : y'_i$
2. $\hat{y} = y \wedge \hat{y'} = y'$
3. $e' \cap c = \emptyset$

The first two conditions are prerequisites for assessment, while the third is the condition which dictates signs of unfaithfulness. We use GPT-4 to insert edits, c instead of GPT-3.5 since this task is much tougher than the previous cases as x' has to correctly correspond to an alternative answer given in the choices while keeping c to a minimal length.

5 Experiments

Datasets: We implement the perturbation experiments across three commonsense reasoning benchmarks.

1. OpenBookQA (Mihaylov et al., 2018), which has 4 answer choices for each question and evaluates open-book reasoning capabilities.
2. QASC (Khot et al., 2020), is an 8-choice multi-hop reasoning dataset requiring assembling multiple real-world facts to successfully answer the question.
3. StrategyQA (Geva et al., 2021) is a binary question dataset structured in a way that the model is required to strategize a chain of reasoning steps to derive the correct answer.

We use only the test set to run the experiments for all perturbations introduced in Section 4, with the exception of LAS, where we employ the LLM to generate explanations for the training set as well.

Model implementation: We use the 70B Llama-v2 (Touvron et al., 2023) from Meta as the choice of LLM for this experiment. In order to fit the implementation within our limited resources, we use a 4-bit quantized version, via applying the GPTQ technique (Frantar et al., 2022), specifically tailored towards GPT-like language models (Brown et al., 2020). The full details of model implementation can be found in Appendix A.2.

Metric details: We use label-flip percentages as the unit of measurement for both paraphrased and mistake insertion. For counterfactual inputs, we only consider an instance as unfaithful if the counterfactual reasoning, e' has a zero overlap with modification c. This applies to most prompting which produces straightforward reasoning explanations, with the exception of QD where we only assessed each sub-answer. We list the prompt templates for generating each perturbation in Appendix A.1. We also compute an overall score, s_o averaging across the four assessments. For paraphrase and counterfactual, we take the complement, $1 - x$, where x is the original unit.

$$s_o = \frac{1}{n} \sum_{i=1}^{n} s_{i,\text{norm}} \quad (1)$$

$$s_{i,\text{norm}} = \frac{s_i - s_{i,\text{min}}}{s_{i,\text{max}} - s_{i,\text{min}}} \quad (2)$$

5.1 Results

We show the full experimental results in Table 1. The proposed approach surpasses all other baseline methods based on the average normalized score, notably displaying a significant difference in OBQA ($> 75\%$). Although SC-CoT is competitive, it still underperforms substantially as compared to SEA-CoT. The other baselines show considerably weaker performance in comparison.

The difference between SC-CoT and SEA-CoT lies with the additional step wherein the model self-critiques its own reasoning, specifically how well the explanation aligns with both the context and the answer it aims to clarify. This has shown significant improvement in both utility as well as minimizing unfaithfulness in counterfactual augmented context. The big leap in the utility scores can mostly be attributed to the fact that having a stimulus aligned with the context, can provide more efficient learning signals to a student model, easing the training process. This can be illustrated in Figure 4, where the word *"shunned"* is mentioned while other baselines used *"would not wear"*, which does not directly relate to the target question, causing the model to erroneously infer the wrong label. While CoT successfully determines the correct answer, it fails to acknowledge the mention of *"Amish cousins"*, thus exhibiting a tenuous connection to the question.

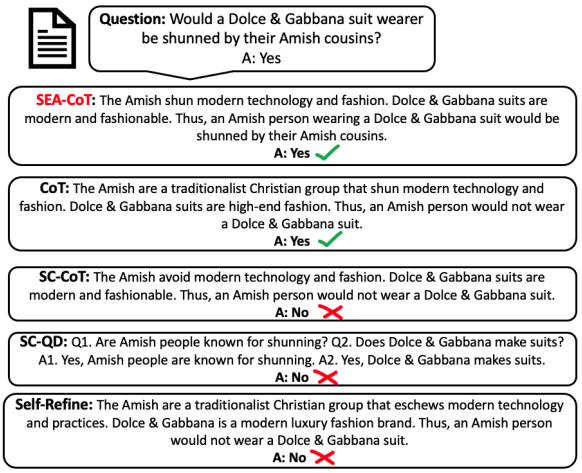

Figure 4: StrategyQA example, the reasoning chain produced by SEA-CoT reflects the important points in the context, making it easier for a learner model to simulate the answer from the given explanation.

A perhaps surprising finding is that Self-Refine performs well below par as compared to the other baselines. Our finding coincides with (Huang et al., 2023), where the authors discuss the flaws of self-correction in reasoning tasks. The primary challenge stems from the intricacy of designing few-shot examples that can effectively drive successive enhancements over prior outputs. Crucially, since the input prompt is already optimized to instruct the LLM for optimal performance in the given task, the potential for self-improvement remains limited. While the SEA-CoT framework requires the LLM to self-assess its responses, it also provides direct guidance aimed specifically at improving a particular attribute: ensuring that the reasoning is consistent with the relevant context. This simple extension greatly improves the quality of the explanation, with no downside on performance.

Dataset/Prompt		Acc (↑)	Robust Para (↓)	Faithful CF-UF(↓)	Faithful Mistakes (↑)	Utility Simu (↑)	Avg (↑)
OBQA	CoT	82.15	3.04	15.54	30.11	17.51	38.94
	QD	79.26	14.01	11.7	23.49	23.93	44.47
	SR	64.55	17.33	**8.6**	**40.82**	10	50
	SC-CoT	83.47	11.27	15.58	37.32	16.05	34.15
	SEA-CoT (Ours)	83.47	**2.45**	9.54	28.78	**30.48**	**67.87**
QASC	CoT	80.26	3.92	20.59	25.5	29.62	41.06
	QD	72.31	20.09	22.22	33.05	30.0	33.27
	SR	65.4	18.36	15.6	**35.74**	18.14	41.94
	SC-CoT	81.57	**2.62**	20.31	30.96	33.89	62.2
	SEA-CoT (Ours)	81.57	2.72	**10.78**	29.43	**38.13**	**84.45**
StrategyQA	CoT	66.92	14.42	7.46	54.06	11.14	46.34
	QD	73.18	14.09	21.26	42.48	2.94	0.62
	SR	66.58	5.09	7.09	**66.69**	8.72	73.24
	SC-CoT	78.51	1.8	6.6	61.8	12.59	82.01
	SEA-CoT (Ours)	78.51	**1.2**	**3.81**	61.24	**16.97**	**94.37**

Table 1: Interpretability results for the 5 prompting techniques across 3 commonsense reasoning benchmark. Three axes of interpretability assessed. 1) Robustness: label flip percentage given paraphrased explanation. 2) Faithfulness: Counterfactual unfaithfulness: instances where modification, c is not reflected in counterfactual explanation, e' and label flip when mistakes are added to explanation. 3) Utility: represented using simulatability of explanation, measured in terms of task enhancement when training context supplemented with explanation. Avg is the combined averaged score across the three axes.

5.2 Few-shot

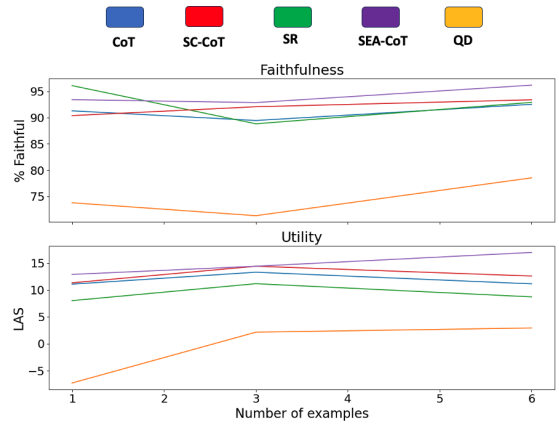

Figure 5: Few-shot performance on both faithfulness (CF-UF) and Utility (LAS) across the five prompts. Assessed on StrategyQA dataset.

We subsequently carry out additional experiments on the number of few-shot examples to study the effects of providing a smaller set of prior examples in the context, displayed in Figure 5. We focus on two qualities: faithfulness and utility. We repeat the same perturbations on StrategyQA across 1,3,6 examples in the input prompt. We choose to assess counterfactual unfaithfulness since assessing mistakes insertion may benefit poorer-performing promptings, given the increased indecisiveness.

A surprising phenomenon can be observed when the LLM produces less faithful reasoning chains when given 3 examples as compared to 6. This is also the case for utility where the performance is not monotonically increasing with the number of examples given, with the exception of SEA-CoT. Notably, even when given a single example, our approach can still outperform the other baselines when the full set of examples is given. This indicates that when the model receives directed feedback, it can more effectively pinpoint areas of focus to produce clearer and more interpretable outputs.

5.3 Model size

Size	Para(↓)	CF-UF (↓)	Mistakes (↑)	Simu (↑)
70B	**1.2**	**3.81**	61.24	**16.97**
13B	4.1	4.38	69.62	6.16
7B	3.79	7.81	**70.62**	15.97

Table 2: Percentage of extracted over target rationales. BoolQ has the lowest percentage out of all three datasets.

The scaling laws of model size primarily concern the downstream performance of LLMs but little is known regarding the influence on interpretability properties. We replicate the experiments on the StrategyQA dataset with a focus on SEA-CoT prompting.

We present the results in Table 2. The largest model, 70B generally outperforms the smaller sizes across all metrics with the exception of having fewer label flips when mistakes are added. This phenomenon might be attributed to the diminished (Acc: 78.51 vs 69.64) performance of smaller mod-

els, which are more susceptible to modifications of their initial decisions, albeit being less robust under similar contexts. Llama-13B surprisingly performs worse than its smaller variant, despite having a bigger network. More importantly, we note that by using SEA-CoT, even a 7B-sized model can generate more interpretable reasoning chains than a 70B model with other baseline prompts.

6 Related Works

Natural Language Explanation (NLE): NLE can primarily be categorized as either abstractive (AE) or extractive (EE). The former is unrestricted by the context and as such enables a higher degree of freedom in explaining the resultant decision. The latter is deemed as the more faithful of the pair as the decision is directly conditioned on the extracted text, though as mentioned earlier should be approached with caution. However, faithfulness is not the only important property, and other properties such as utility and plausibility should be present to ensure sufficient interpretability. EE typically falls short in the realm of plausibility since humans do not understand spans of text without a full context in view (Gurrapu et al., 2023). In this work, the subject of interest is in AE.

Given the advance in text generation models, researchers are leaning towards AE in hopes of producing explanations that can be easily understood by the layperson. (Majumder et al., 2021) utilizes a union of both forms of explanation, conditioning the generation of AE on the extracted spans of text while concurrently grounding the generation on relevant world knowledge. The resultant interpretation is then assumed to be faithful while plausible. Similar works include faithfulness through task decomposition (Sanyal et al., 2022), label-specific explanations (Kumar and Talukdar, 2020). (Narang et al., 2020) demonstrate the possibility of inducing plausible explanations simply by pretending the word *explain* to the input prompt, similar to how CoT works.

Interpretable CoT: Since its introduction, CoT has seen widespread usage due to its simplicity and intuition it offers and has garnered interest in the research community to innovate adaptation of it (Chu et al., 2023). Despite CoT being primarily introduced to facilitate better reasoning skills out of LLMs, there is much interest to see if these reasoning steps could be used as a form of explaining the model's thought process. Most of such works primarily investigate the faithfulness of the reasoning (Lanham et al., 2023; Radhakrishnan et al., 2023; Turpin et al., 2023) or improving the faithfulness in CoT outputs, via refinement through knowledge retrieval (He et al., 2022), symbolic reasoning (Lyu et al., 2023), iterative information selection (Creswell and Shanahan, 2022) and factuality calibration (Ye and Durrett, 2022).

Concurrently, other works (Wang et al., 2023; He et al., 2022) are focused on ascertaining the faithfulness of an explanation to the presence of factuality. While factuality is an important trait, it is not a sufficient component to represent faithfulness. Non-factual explanations may still align faithfully with an incorrect answer, as long as the explanation is aligned with the incorrect label inferred. Our work strives to conduct a holistic assessment of interpretability across various forms of prompting techniques used in LLMs, taking into account multiple important properties which may be of importance towards various audiences.

7 Conclusion

This work introduces multiple ways to assess the interpretability of an explanation. The focus of this work is centered around the different variants of CoT and how we can better determine the usability of the reasoning by-product as an explanation for the underlying prediction. We also propose a modification to the SC-CoT framework called *SEA-CoT*, designed specifically to yield explanations that better fulfill the objectives of interpretability. Our proposed framework surpasses the Robustness, Faithfulness, and Utility dimensions across multiple reasoning benchmarks. In the future, we plan to extend our work towards instilling interpretability and safety in the training stages (Yang et al., 2023), such as safety alignment in LLM.

8 Limitations

Our work only investigates a single LLM - Llama-2 This work could be extended toward transformers of different structures such as encoder or encoder-decoder, or larger models, such as GPT3.5/4.0, which due to limiting resources are restricted to generate assessments instead. This work left out other techniques such as grounding the LLM's response via external knowledge, which we note is an interesting avenue to consider next.

References

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. 2023. Faithfulness tests for natural language explanations. *arXiv preprint arXiv:2305.18029*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

Howard Chen, Jacqueline He, Karthik Narasimhan, and Danqi Chen. 2022. Can rationalization improve robustness? *arXiv preprint arXiv:2204.11790*.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2023. A survey of chain of thought reasoning: Advances, frontiers and future. *arXiv preprint arXiv:2309.15402*.

Antonia Creswell and Murray Shanahan. 2022. Faithful reasoning using large language models. *arXiv preprint arXiv:2208.14271*.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2022. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*.

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.

Sai Gurrapu, Ajay Kulkarni, Lifu Huang, Ismini Lourentzou, Laura Freeman, and Feras A Batarseh. 2023. Rationalization for explainable nlp: A survey. *arXiv preprint arXiv:2301.08912*.

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? *arXiv preprint arXiv:2010.04119*.

Hangfeng He, Hongming Zhang, and Dan Roth. 2022. Rethinking with retrieval: Faithful large language model inference. *arXiv preprint arXiv:2301.00303*.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? *arXiv preprint arXiv:2004.03685*.

Alon Jacovi and Yoav Goldberg. 2021. Aligning faithful interpretations with their social attribution. *Transactions of the Association for Computational Linguistics*, 9:294–310.

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.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Sawan Kumar and Partha Talukdar. 2020. Nile: Natural language inference with faithful natural language explanations. *arXiv preprint arXiv:2005.12116*.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. *arXiv preprint arXiv:1606.04155*.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-ofthought 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, et al. 2023. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.

Bodhisattwa Prasad Majumder, Oana-Maria Camburu, Thomas Lukasiewicz, and Julian McAuley. 2021. Knowledge-grounded self-rationalization via extractive and natural language explanations. *arXiv preprint arXiv:2106.13876*.

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. *arXiv preprint arXiv:1809.02789*.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. Wt5?! training text-to-text models to explain their predictions. *arXiv preprint arXiv:2004.14546*.

Liangming Pan, Michael 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 preprint arXiv:2308.03188*.

Ansh Radhakrishnan, Karina Nguyen, Anna Chen, Carol Chen, Carson Denison, Danny Hernandez, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, et al. 2023. Question decomposition improves the faithfulness of model-generated reasoning. *arXiv preprint arXiv:2307.11768*.

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.

Soumya Sanyal, Harman Singh, and Xiang Ren. 2022. Fairr: Faithful and robust deductive reasoning over natural language. *arXiv preprint arXiv:2203.10261*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *arXiv preprint arXiv:2305.04388*.

Keheng Wang, Feiyu Duan, Sirui Wang, Peiguang Li, Yunsen Xian, Chuantao Yin, Wenge Rong, and Zhang Xiong. 2023. Knowledge-driven cot: Exploring faithful reasoning in llms for knowledge-intensive question answering. *arXiv preprint arXiv:2308.13259*.

Peifeng Wang, Aaron Chan, Filip Ilievski, Muhao Chen, and Xiang Ren. 2022a. Pinto: Faithful language reasoning using prompt-generated rationales. *arXiv preprint arXiv:2211.01562*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022b. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Sarah Wiegreffe, Ana Marasović, and Noah A Smith. 2020. Measuring association between labels and free-text rationales. *arXiv preprint arXiv:2010.12762*.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. 2023. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *arXiv preprint arXiv:2304.13712*.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. *Advances in neural information processing systems*, 35:30378–30392.

Wei Jie Yeo, Wihan van der Heever, Rui Mao, Erik Cambria, Ranjan Satapathy, and Gianmarco Mengaldo. 2023. A comprehensive review on financial explainable ai. *arXiv preprint arXiv:2309.11960*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

A Appendix

A.1 Perturbation details

We use GPT3.5 to generate paraphrased versions of the reasoning explanation produced by prompting the LLM, with the exception of QD. For QD, we select one subquestion-answer pair to apply the perturbations to, we paraphrase both chosen question-answer pairs and only add mistakes to the answer as the focus is on producing wrong answers and not incomprehensible questions. To convert the question x to a counterfactual instance x', we use GPT4 as GPT3.5 frequently produces nonsensical questions that the available answer options cannot answer. Furthermore, we subsequently deploy GPT3.5 again to identify the edited and original portions of x, namely the modification c. Thus, we end up with two sets of templates for both paraphrasing and addition of mistakes (one for QD, one for others) and one set of counterfactual generation. We use 2-shot examples for adding mistakes, 3-shot for counterfactual generation, and 0-shot for paraphrasing. All figures are from Figure 6 to 10

A.2 Inference details

As we do not use API for the bulk of the experiments with the exception of perturbation generation and ablation using GPT3-5. We mainly rely on local resources to conduct inference. We use 4 x A6000 GPU for all experiments, each GPU has 46GB of VRAM and this gives us a total of 184GB VRAM. A 70B model would require at least 140GB VRAM, leaving only 44 VRAM left for text generation. Given an average input size of 1000 (usually longer for prompts such as QD) and a single batch size of 1, it would require an additional >60 GB VRAM (computed based on L = 80, H= 64, dim = 8192 for 70B) which makes it infeasible to implement. Thus, we perform the experiments using a 4-bit quantized version instead, which is performed using GPTQ on the original Llama-2 70B model. GPTQ has been shown to be suitable for quantizing models consisting of billions of parameters. It has been validated on models up to 176B parameters and shown comparable performance with 16-bit models. The GPTQ-ed models are readily available on `huggingface`.

We utilized `text-generation-inference`, an optimized platform for conducting fast inference on LLMs by huggingface, to speed up the inference process. Overall, this allows us to process up to a batch size of 16 across the full hardware stack.

A.3 Few-shot Prompts

We show the few-shot examples used for OBQA dataset, highlighting the differences in the instruction prompt between the various techniques reviewed. The few-shot examples are similar to (Wei et al., 2022), and adjusted when neccessary, depending on the specific prompting methodology.

For Self-Refine, there are three stages of instruction-prompting, where the second (feedback) and third (refine) stages continue iteratively until the LLM detects a stopping criterion which ends the cycle, denoted as *"Stop refining the answer."*. In the *initial generation*, the optimal examples are given, similar to CoT. In the *feedback* stage, we list scoring criteria which is focused on improving the interpretability of the reasoning explanation, instead of focusing on the performance. To simulate various qualities of output, we include both positive and negative examples. The examples in the refine stage are similar to the feedback but are instead designed in a continuous conversion displaying the full process of refining a bad example into a good one. We limit the number of examples in the *refine* stage to 3 as the context length is much longer here. The few-shot example prompts are displayed from Figure 11 to 14.

A.4 Entailment Generation

We designed a separate prompt to be used solely by SEA-CoT, where the LLM is instructed to self-critique the entailment between its own reasoning chain and the combined context of both the question and the produced answer. We use samples from the e-SNLI dataset (Camburu et al., 2018), we only picked instances corresponding to either entailment or contradiction and left out the neutral ones, as the LLM is only instructed to infer if the explanation entails or contradicts the target context.

The probabilities for the entailment label *"yes"* is directly used while we take the complement if generated *"no"*, with the assumption that other tokens in the vocabulary are negligible. The examples are displayed in Figure 15.

0-shot Paraphrasing

User: Please rewrite the following text, conveying exactly the same information but using different wording
Text: A full bag swung in circles is an example of force causing an increase in speed.

Assistant: An instance of force leading to an acceleration in velocity can be observed when a bag, filled to capacity, rotates in circular motion.

0-shot Paraphrasing (QD)

User: Please rewrite the following question and answer, conveying exactly the same information but using different wording.
Question: What is the relationship between force and speed?
Answer: Force can cause an increase in speed.

Assistant:
Paraphrased Question: How are force and speed related?
Paraphrased Answer: Force has the ability to facilitate an acceleration in speed.

Figure 6: 0-shot paraphrase template. Input [Underline] Generated: [highlighted]

2-shot Add Mistakes

User: Given a question and a sentence of reasoning that was used to help answer that question. Please give me a new version of that reasoning sentence, but with at least one mistake added that will change the sentence from factual to fictional.
Question: The sun is responsible for,
Choices:
(a) puppies learning new tricks
(b) children growing up and getting old
(c) flowers wilting in a vase
(d) plants sprouting, blooming and wilting
Reasoning: A plant requires sunlight for photosynthesis, which accumulates resources required for sprouting, blooming and wilting.
Assistant: Reasoning with mistakes: A plant requires money to grow, which requires accumulating resources.
...

User: Given a question and a sentence of reasoning that was used to help answer that question. Please give me a new version of that reasoning sentence, but with at least one mistake added that will change the sentence from factual to fictional.
Question: Which best demonstrates the concept of force causing an increase in speed?
Choices:
(a) skating on a rough surface
(b) a full bag swung in circles
(c) a computer powering on
(d) a baker stirring batter
Reasoning: A full bag swung in circles is an example of force causing an increase in speed.
Assistant: A full bag swung in circles is an example of force causing a decrease in speed.

Figure 7: 2-shot inserting mistake template for all prompt except QD. Input [Underline] Generated: [highlighted]. Only show 1 example.

2-shot Add Mistakes (QD)

User: Given a question and an answer. Please give me a new version of the answer with at least one mistake added, such that it changes from factual to fictional and is no longer the correct answer.
Question: The sun is responsible for,
Answer: plants sprouting, blooming and wilting
Assistant: plants dying due to lack of sunlight.
...

User: Given a question and an answer. Please give me a new version of the answer with at least one mistake added, such that it changes from factual to fictional and is no longer the correct answer.
Question: What is the relationship between force and speed?
Answer: Force can cause an increase in speed.
Assistant: Answer with mistakes: Force can cause a decrease in speed.

Figure 8: 2-shot inserting mistake template for QD. Input [Underline] Generated: [highlighted]. Only show 1 example.

3-shot Counterfactual generation

User: Given a question and corresponding answer, please choose the second most likely answer from the answer choices and generate a new question such that the new question will correspond to the second likely answer. You are to make minimal changes to the question.
Question: Which of these things will supplement a squirrel's caloric needs?
Answer: (c) nuts
Answer Choices:
(a) fish
(b) Tree bark
(c) nuts
(d) Leaves
Assistant: The second most likely answer is: (a) fish
Counterfactual question: Which of these things will supplement a bear's caloric needs?

...

User: Given a question and corresponding answer, please choose the second most likely answer from the answer choices and generate a new question such that the new question will correspond to the second likely answer. You are to make minimal changes to the question.
Question: Which best demonstrates the concept of force causing an increase in speed?
Answer: a full bag swung in circles
Answer Choices:
(a) skating on a rough surface
(b) a full bag swung in circles
(c) a computer powering on
(d) a baker stirring batter
Assistant: The second most likely answer is: skating on a rough surface
Counterfactual question: Which best demonstrates the concept of friction causing a decrease in speed?

Figure 9: 3-shot counterfactual generation Input [Underline] Generated: [highlighted]. Only show 1 example. First, identify the next possible answer before editing the question towards it.

0-shot Edit highlighting
User: Please highlight the difference in text between the original statement and changed statement. Only highlight the extra or changed words and ignore the missing ones. If there are more than one contiguous text, use a comma to separate them.
Original statement: Which best demonstrates the concept of force causing an increase in speed?
Changed statement: Which best demonstrates the concept of friction causing a decrease in speed?

Assistant:
Original: force causing an increase in speed
Changed: friction causing a decrease in speed

Figure 10: 0-shot edit highlighting. Input [Underline] Generated: [highlighted]. Identify edits corresponding to original text.

Prompt for OBQA (CoT, SC-CoT, SEA-CoT)

Q: The sun is responsible for what?
Answer choices: (a) Puppies learning new tricks (b) Children growing up and getting old (c) Flowers wilting in a vase (d) Plants sprouting, blooming and wilting
A: A plant requires sunlight for photosynthesis, which accumulates resources required for sprouting, blooming and wilting. So the answer is (d)

Q: When standing miles away from Mount Rushmore, what happens?
Answer choices: (a) The mountains seem very close (b) The mountains are boring (c) The mountains look the same as from up close (d) The mountains seem smaller than in photographs
A: When an object is far away, it takes up less of your field of view, and so seems smaller than in the photographs. So the answer is (d)

Q: When food is reduced in the stomach, what happens?
Answer choices: (a) The mind needs time to digest (b) Take a second to digest what I said (c) Nutrients are being deconstructed (d) Reader's digest is a body of works
A: The stomach is part of the digestive system. The breaking down of food into nutrients occurs in the digestive system. So the answer is (c)

Q: Poison causes harm to which of the following?
Answer choices: (a) A Tree (b) A robot (c) A house (d) A car
A: A tree is a living thing. Poison causes harm to living things. So the answer is (a)

Q: A magnet will stick to what?
Answer choices: (a) A belt buckle (b) A wooden table (c) A plastic cup (d) A paper plate
A: A belt buckle is made of metal. If a magnet is attracted to a metal then that magnet will stick to that metal. So the answer is (a)

Q: Deer are less safe in the woods because wolves have what?
Answer choices: (a) Have fur (b) Howl (c) Have claws (d) Have tails
A: Claws are used by wolves to catch prey like deer. So the answer is (c)

Q: An electric car causes what?
Answer choices: (a) More CO2 emissions (b) Equal CO2 emissions (c) Electric emissions (d) Less CO2 emissions
A: An electric car uses less gasoline than a regular car and thus causes less CO2 emissions. So the answer is (d)

Figure 11: 7-shot prompt used for CoT, SC-CoT and SEA-CoT. There are newlines between answer choices and each given choice, opted out to save space.

Prompt for OBQA (QD)

Q: The sun is responsible for what?
Answer choices: (a) Puppies learning new tricks (b) Children growing up and getting old (c) Flowers wilting in a vase (d) Plants sprouting, blooming and wilting
Lets break down the problem.
Q1. What do the sun provide?? Q2. What can sunlight be used for? A1. The sun provides sunlight. A2. Sunlight can be used for photosynthesis which accumulates resources required for sprouting, blooming, and wilting of plants. Thus, sunlight is responsible for plants sprouting, blooming, and wilting. So the answer is (d)

Q: When standing miles away from Mount Rushmore, what happens?
Answer choices: (a) The mountains seem very close (b) The mountains are boring (c) The mountains look the same as from up close (d) The mountains seem smaller than in photographs
Lets break down the problem.
Q1. What happens to the perception of an object when it's far away? Q2. How does an object's size in the field of view relate to its perceived size? A1. When an object is far away, it takes up less of your field of view. A2. An object seems smaller when it takes up less of your field of view than in photographs. Thus, when standing miles away from Mount Rushmore, the mountains seem smaller than in photographs. So the answer is (d)

Q: When food is reduced in the stomach, what happens?
Answer choices: (a) The mind needs time to digest (b) Take a second to digest what I said (c) Nutrients are being deconstructed (d) Reader's digest is a body of works
Lets break down the problem.
Q1. Where does the breaking down of food into nutrients occur? Q2. What is the role of the stomach in digestion? A1. The breaking down of food into nutrients occurs in the digestive system. A2. The stomach is part of the digestive system. Thus, when food is reduced in the stomach, nutrients are being deconstructed. So the answer is (c)

Q: Poison causes harm to which of the following?
Answer choices: (a) A Tree (b) A robot (c) A house (d) A car
Lets break down the problem.
Q1. What does poison harm? Q2. Which of the choices are living things? A1. Poison causes harm to living things. A2. A tree is a living thing. Thus, poison causes harm to a tree. So the answer is (a)

Figure 12: 7-shot prompt used for QD. We show only 4 examples here, and there are newlines between each sub-questions and answers, which we similarly leave out to save space.

Prompt for OBQA (Self-Refine)

Initial generation:
Given a question with answer choices, generate a reasoning explanation which supports the selected answer. Desired traits for the reasoning are explanation are 1) Factuality - The reasoning should be factual and should not contain any false information. 2) Relevance - The reasoning should be relevant to both the question and answer. 3) Informativeness - The reasoning should provide sufficient information to support the answer.

Q: The sun is responsible for what?
Answer choices: (a) Puppies learning new tricks (b) Children growing up and getting old (c) Flowers wilting in a vase (d) Plants sprouting, blooming and wilting
A: A plant requires sunlight for photosynthesis, which accumulates resources required for sprouting, blooming and wilting. So the answer is (d)

Feedback
We want to iteratively improve the provided responses. To help improve, scores for each response on desired traits are provided: 1) Factuality, 2) Relevance, 3) Informativeness. Please rate each trait from 1 to 5 and decide if the answer requires further refinement. If not, append 'stop refining the answer' to the end of the feedback.

Q: The sun is responsible for what?
Answer choices: (a) Puppies learning new tricks (b) Children growing up and getting old (c) Flowers wilting in a vase (d) Plants sprouting, blooming and wilting
A: A plant requires sunlight for photosynthesis, which accumulates resources required for sprouting, blooming and wilting. So the answer is (d)
Scores:
Factuality: Sentence is factual. 5/5. Relevance: Sentence is directly relevant to the question. 5/5. Informativeness: The sentence provides accurate reasoning. 5/5
Total Score: 15/15. Stop refining the answer.

Q: When standing miles away from Mount Rushmore, what happens?
Answer choices: (a) The mountains seem very close (b) The mountains are boring (c) The mountains look the same as from up close (d) The mountains seem smaller than in photographs
A: Mount Rushmore is a mountain. Mountains always seem closer when viewed from far away. So the answer is (a).
Scores:
Factuality: Incorrect logic about mountains. 1/5. Relevance: The statement is not entirely relevant to the question. 2/5. Informativeness: Provides a misguided reasoning. 2/5
Total Score: 5/15

Figure 13: Prompt for Self-Refine, we show a single example for the initial generation, the rest is similar in CoT examples. For the feedback, we include both good and bad examples, both displayed here. We use 7 examples for both initial generation and feedback.

Refine:
We want to iteratively improve the provided responses. To help improve, scores for each response on desired traits are provided: 1) Factuality, 2) Relevance, 3) Informativeness. Please rate each trait from 1 to 5 and decide if the answer requires further refinement. If not, append 'stop refining the answer' to the end of the feedback.

Q: When standing miles away from Mount Rushmore, what happens?
Answer choices: (a) The mountains seem very close (b) The mountains are boring (c) The mountains look the same as from up close (d) The mountains seem smaller than in photographs
A: Mount Rushmore is a mountain. Mountains always seem closer when viewed from far away. So the answer is (a).
Scores:
Factuality: Incorrect logic about mountains. 1/5. Relevance: The statement is not entirely relevant to the question. 2/5. Informativeness: Provides a misguided reasoning. 2/5
Total Score: 5/15

Okay, improve the sentence using the feedback:

A: Mountains, in general, can be less detailed from far away, but their size perception doesn't change. So the answer is (c).
Scores:
Factuality: Inaccurate understanding of perspective. 2/5. Relevance: It somewhat addresses the question about viewing the mountain from a distance. 3/5. Informativeness: The sentence is somewhat informative. 3/5
Total Score: 8/15

Okay, improve the sentence using the feedback:

A: When an object is far away, it takes up less of your field of view, and so seems smaller than in the photographs. So the answer is (d)
Scores:
Factuality: Sentence is factual. 5/5. Relevance: Sentence is directly relevant to the question. 5/5. Informativeness: The sentence provides accurate reasoning. 5/5
Total Score: 15/15. Stop refining the answer.

Figure 14: Refine stage in Self-Refine, we show a single example here, where each example demonstrates the entire refining process from a bad to good example.

Self-Entailment-Alignment (entailment generation)

Given a premise and hypothesis, predict if the hypothesis entails the premise.
Premise: This church choir sings to the masses as they sing joyous songs from the book at a church. Hypothesis: The church is filled with song.
A: yes

Given a premise and hypothesis, predict if the hypothesis entails the premise.
Premise: This church choir sings to the masses as they sing joyous songs from the book at a church. Hypothesis: A choir singing at a baseball game.
A: no

Given a premise and hypothesis, predict if the hypothesis entails the premise.
Premise: A woman with a green headscarf, blue shirt and a very big grin. Hypothesis: The woman is very happy.
A: yes

Given a premise and hypothesis, predict if the hypothesis entails the premise.
Premise: An old man with a package poses in front of an advertisement. Hypothesis: A man walks by an ad.
A: no

Given a premise and hypothesis, predict if the hypothesis entails the premise.
Premise: A statue at a museum that no seems to be looking at. Hypothesis: The statue is offensive and people are mad that it is on display.
A: no

Given a premise and hypothesis, predict if the hypothesis entails the premise.
Premise: A land rover is being driven across a river. Hypothesis: A Land Rover is splashing water as it crosses a river.
A: yes

Given a premise and hypothesis, predict if the hypothesis entails the premise.
Premise: A man playing an electric guitar on stage. Hypothesis: A man playing guitar on stage.
A: yes

Figure 15: NLI examples for entailment generation for SEA-CoT, used across all datasets.