How Interpretable are Reasoning Explanations from Prompting Large Language Models?

Wei Jie Yeo[1], **Ranjan Satapathy**[2], **Rick Siow Mong Goh**[2], **Erik Cambria**[1]
[1]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[2]Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A∗ STAR),
1 Fusionopolis Way, #16-16 Connexis, 138632, Singapore
[1]`yeow0082@e.ntu.edu.sg`,
[2]`{satapathy_ranjan,gohsm}@ihpc.astar.edu.sg`,
[1]`cambria@ntu.edu.sg`

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 across multiple commonsense reasoning benchmarks. 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 interpretability alignment technique, termed Self-Entailment-Alignment Chain-of-thought, that yields more than 70% improvements across multiple dimensions of interpretability. Code is available at `https://github.com/SenticNet/CoT_interpretability`

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 attributing factor lies with the extensive efforts put into innovating new ways of prompting the LLM to better exploit their knowledge base. 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), Least-to-Most (Zhou et al., 2022), while others specifically focus on establishing faithful reasoning (Creswell and Shanahan, 2022; Lyu et al., 2023). 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)* operates similarly to Self-Consistency, but additionally utilizes a form of consistency between the corresponding reasoning steps and supported context. This action is missing in Self-Consistency, as the focus is only on the resultant output, potentially leading to unfaithful reasoning which may not support the underlying answer.

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

Findings of the Association for Computational Linguistics: NAACL 2024, pages 2148–2164
June 16-21, 2024 ©2024 Association for Computational Linguistics

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 authors mentioned 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 hypothesize that besides the limitation of narrowing the predictors' context, generating the explanation and output using separate modules could compromise the quality of the explanation. We set up a simple study, comparing a modular against a single LLM setup on two interpretability traits, faithfulness, and utility, covered in deeper detail in section 5. We adopt the PINTO framework (Wang et al., 2022a), where the explainer, r_θ is a frozen pre-trained LLM and the predictor, f_ϕ is finetuned on the downstream task, conditioned on both the generated explanation and context, $\hat{y} = f_\phi(x \oplus \hat{e})$, where $\hat{e} = r_\theta(x)$, x is the given context and $\oplus$ is the concatenation process.

For the single LLM setup, we directly 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, $e_{\hat{j \neq i}}$ 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 $\hat{y}$. The details of LAS are covered in A.7

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 being given an unrelated stimulus.

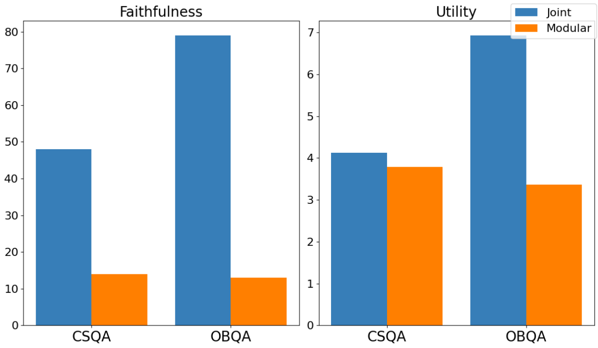


Figure 1: Faithfulness and Utility scores for joint and modular approach on two reasoning datasets: CSQA and OBQA.

Notably, this observation resonates well with the recognized capability of recent LLMs to self-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. This preliminary experiment serves as the main motivation to conduct experiments to scrutinize the quality of explanations produced by a singular LLM.

3 Prompt Techniques

In this section, we systematically review various ways an LLM, f_ϕ can be prompted. These methods primarily differ in how the language model is queried to derive the final answer. Furthermore, we proposed an approach, SEA-CoT, aimed at improving the interpretability traits of the reasoning chain to serve as the explanation for the resultant output. 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 greedy decoding in CoT by sampling multiple, N paths and choosing the final answer,

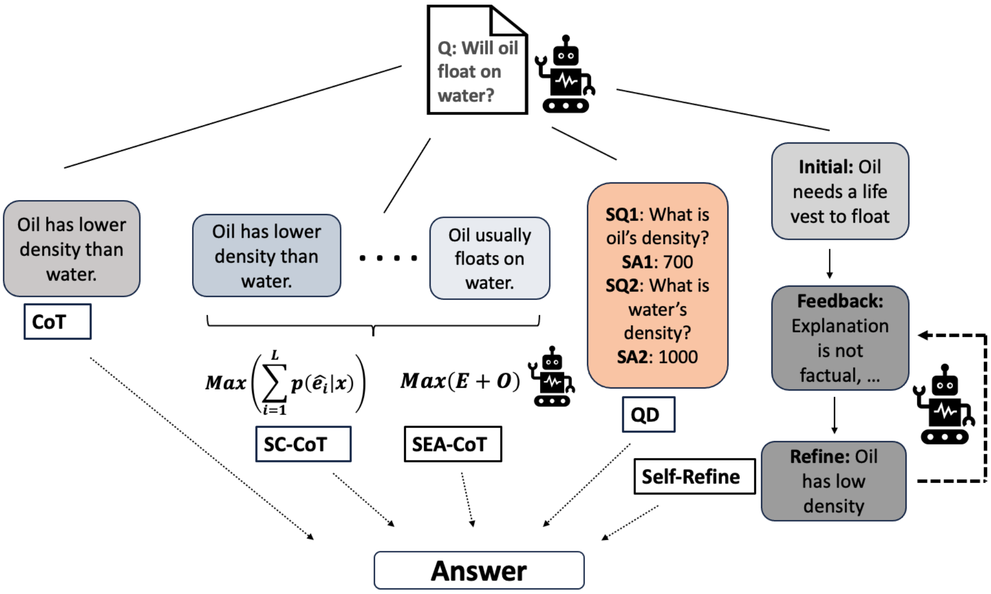

Q: Will oil float on water?
Oil has lower density than water.
CoT
Oil has lower density than water.
Oil usually floats on water.
$Max\left(\sum_{i=1}^{L} p(\hat{e}_i|x)\right)$
$Max(E+O)$
SC-CoT
SEA-CoT
SQ1: What is oil's density?
SA1: 700
SQ2: What is water's density?
SA2: 1000
QD
Initial: Oil needs a life vest to float
Feedback: Explanation is not factual, ...
Refine: Oil has low density
Self-Refine
Answer

Figure 2: Overview of different prompting techniques to derive the reasoning chain, to serve as the explanation (boxed with dashed line). [Left to Right]: Cot, SC-CoT, SEA-CoT, QD, and Self-Refine (SR). SC-CoT and SEC-CoT differ in the explanation selection stage, where the former selects based on maximum cumulative probability and the latter 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 a stopping criteria is encountered or exceeds the max number of passes. SR requires the most pass, 3 per round.

$\hat{y}^*$ via majority voting. SC-CoT has shown improvements across multiple arithmetics and commonsense reasoning benchmarks. Since multiple explanations may lead to the majority answer. We choose the explanation with the highest cumulative probability. We also experiment with different ways, further discussed in the ablation section.

- **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 conclude 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.

4 Proposed Approach

Most adaptations on CoT are only aimed at maximizing task performance as covered in Section 3. Our work is instead focused on enhancing the interpretability of the presented reasoning chain preceding the task output. We adapt from SC-CoT, by focusing on the N sequences produced, ranked based on specified objectives. Instead of picking explanations based on heuristics such as highest cumulative probability, the reasoning is chosen based on the maximization of two objectives: entailment and overlapping score between the supported context $(x \oplus \hat{y})$ and reasoning $\hat{e}$.

We posit that a credible explanation should intrinsically align with the given context it aims to elucidate (Jie et al., 2024); in this scenario, it encompasses both the question being addressed and the prediction label, measured by the level of entailment.

We additionally maximize the overlap between two sets of key tokens[1], which we show in later experiments to be beneficial towards producing higher quality explanations. This simple approach can be regarded as performing a self-alignment step to pick the most suitable explanation with the N sequences.

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 each own generated reasoning, $\hat{e}_i$ and the joint context, $x \oplus \hat{y}$. We prompt the LLM with few-shot examples of natural language inference (NLI), x_e in Figure 16 and determine if the hypothesis entails the premise. The final score to be ranked, S_T is a combination of both the probability of entailment, S_e, and the IoU score, S_o.

$$S_e = p_e(f_\phi(x \oplus \hat{y}, \hat{e}_i | x_e)) \tag{1}$$

$$S_o = \frac{|\hat{e}_i \cap (x \oplus \hat{y})|}{|\hat{e}_i \cup (x \oplus \hat{y})|} \tag{2}$$

$$S_T = S_e + S_o \tag{3}$$

The most interpretable explanation is then chosen via maximizing S_T. One caveat is that in the event where $|\hat{y}^*| = 1$, we fall back to SC-CoT. However, this can be avoided by trivially setting the number of sequences, N to be higher than the number of possible options.

5 Interpretability Qualities

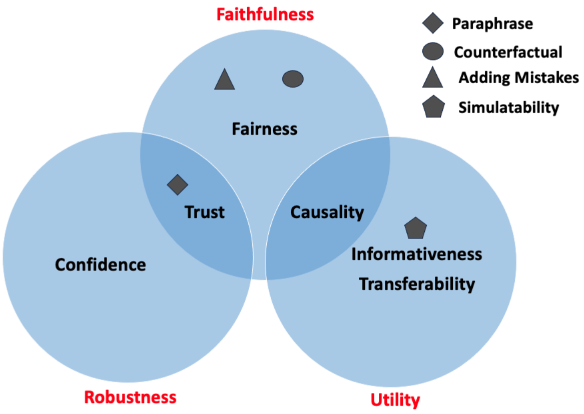


Figure 3: The interpretability qualities measured by different perturbation tests, to achieve the corresponding goals of an explanation. Goals referenced from (Yeo et al., 2023)

Interpretability is a multifaceted characteristic with multiple desirable traits concerning various goals of interpretability. Inspired by existing work on desirable goals of explainable AI (Yeo et al., 2023), we assess three aspects of interpretability: faithfulness, robustness, and utility. We propose these traits as we believe they are directly linked to achieving such goals, illustrated in Figure 3. We discuss the connections in further sections. In accordance, we outline the corresponding evaluations sought out to assess each trait, shown in Figure 4. These evaluations are primarily conditioned on both the context and self-generated reasoning chain.

Faithfulness: 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 tests assessing the faithfulness of reasoning chains generated using CoT from an LLM. However, the authors only investigated a single prompting technique, while we conducted extensive experiments covering multiple prompting approaches. A faithful explanation is crucial as it fosters trust (Cambria et al., 2023) and fairness, ensuring that users can rely on the explanation to reflect the decision-making process and identify any potential biases, thereby improving model transparency and understanding of any causal relationships

Robustness: Robustness 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. A robust explanation instills confidence and trust in users that the model would behave appropriately despite noises in the input.

Utility: A largely understudied but important trait, utility is paramount to maximizing the information conveyed to the audience. A useful explanation can allow the discovery of new knowledge to human users such as understanding the causal relationships or enable more efficient knowledge distillation between neural models.

---

[1] The two sets of tokens are compared after removing any stopwords to minimize noise within the context

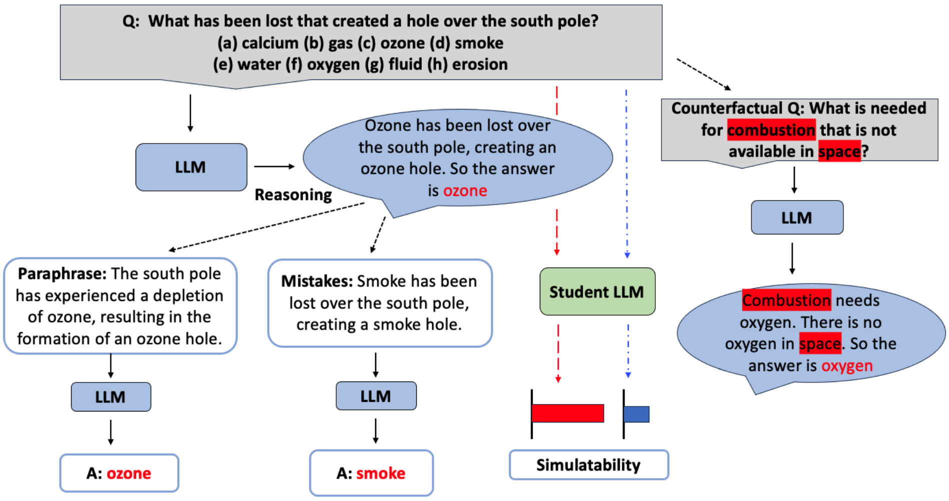

Q: What has been lost that created a hole over the south pole?
(a) calcium (b) gas (c) ozone (d) smoke
(e) water (f) oxygen (g) fluid (h) erosion
LLM
Reasoning
Ozone has been lost over the south pole, creating an ozone hole. So the answer is ozone
Paraphrase: The south pole has experienced a depletion of ozone, resulting in the formation of an ozone hole.
LLM
A: ozone
Mistakes: Smoke has been lost over the south pole, creating a smoke hole.
LLM
A: smoke
Student LLM
Simulatability
Counterfactual Q: What is needed for combustion that is not available in space?
LLM
Combustion needs oxygen. There is no oxygen in space. So the answer is oxygen

Figure 4: Interpretability test for faithfulness, robustness, and utility. The reasoning chain is subjected to perturbations: paraphrasing and inserting mistakes, before re-generating the subsequent output. Counterfactual: the original question is changed to check if the resultant reasoning accounts for edits (shaded red). Simulatability: increase in task performance when training data is augmented with reasoning chain, measured with a student model.

5.1 Paraphrase

Paraphrasing $\hat{e}$ corresponding to $\hat{y}$ allows us to question the robustness of the explanation, ie how robust is the explanation against minor variations, assuming that these variations do not alter the core intent, yet still enable the model to produce the same outcome? Albeit such a test concurrently touches on the concept of faithfulness, where similar thought processes should lead to identical conclusions given the same model (Jacovi and Goldberg, 2020). However, for the sake of differentiation, we consider the primary objective of paraphrasing as an evaluation of robustness in the following experiments.

5.2 Adding mistakes

In contrast to ensuring answer consistency among similar reasoning, inserting erroneous inputs into an explanation can assess if the reasoning preceding the output is truly faithful. One would expect the model to change its decision given an erroneous reasoning chain if it is faithful from the start. We focus on the alteration in prediction rather than actual task performance, since incorrect reasoning may potentially correct an erroneous explanation, though such occurrences are exceedingly rare.

5.3 Simulatability

As it is costly to employ humans to assess if a reasoning chain is useful, we employ forward simulatability as a proxy for utility. We measure simulatability using LAS in Section 2 (further details in A.7) as it highly correlates with human judgment. A 220M T5-base (Raffel et al., 2020) is selected as the student model, to measure utility from improvement in downstream performance. 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 explanations, followed by an evaluation of its performance. A key aspect of LAS lies with the notion of subtracting a baseline, $M_s(f_s(x))$ from $M_s(f_s(\hat{e} \oplus x))$, where M_s is a task scoring function such as accuracy or F1-score. This is used to assess the benefits gained by adding $\hat{e}$ into the training process.

5.4 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, $y' \neq y$. The distinction from Section 5.2 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 can be seen as an undirected measure aimed at gauging the decline in confidence, given an erroneous prior belief. We deemed an instance of unfaithfulness 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 indicates signs of unfaithfulness.

6 Experiments

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

1. OpenBookQA (Mihaylov et al., 2018), which has 4 answer choices for each question and assesses 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 such 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 5, except LAS, where we employ the LLM to generate explanations for the training set as well.

Implementation details: We use the 70B Llama-v2 (Touvron et al., 2023) from Meta as the choice of LLM for this experiment. We use a 4-bit quantized version, via applying the GPTQ technique (Frantar et al., 2022) since the full 32-bit model would require extensive resources. The full model implementation details can be found in Appendix A.2.

Explanation modifications: We perform automatic checks on the modifications to prevent errors in the experiments. In paraphrase, the modified explanation is chosen only when the resultant output $\hat{y}|e_m$ remains the same, and the opposite for mistakes insertion. In counterfactual generation. OpenAI's GPT3.5 is used for both paraphrasing and mistake insertion and GPT4 for counterfactuals since the task is much harder as x' has to correspond to an alternate answer choice.

Metric details: We use the percentage of flipped predictions as the measurement unit for both paraphrased and mistake insertion. For counterfactual inputs, we only consider e to be unfaithful if the counterfactual part, e' has a zero overlap with modification c. This applies to singular reasoning chains, except QD where we only assessed each sub-answer. Utility is measured using the LAS score, corresponding to the increase in performance when supplemented with explanation during training. We list the prompt templates in Appendix A.1. We compute an aggregate score, averaging across the four qualities, after normalizing each score between 0 and 1. For paraphrase and counterfactual, we take the complement score, $1 - s$, where s is the original unit.

6.1 Results

We show the full experimental results in Figure 5. SEA-CoT surpasses all other baseline methods based on the average normalized score, notably displaying a significant difference in OBQA ($> 75\%$) over majority of the baselines. Although SC-CoT is competitive, it still underperforms substantially as compared to SEA-CoT. We observe that the underperformance of SEA-CoT in the mistakes criteria can be explained via the relationship between SR's weak task performance and high score in mistakes, attributed to a higher likelihood of altering its output. Whereas, SEA-CoT achieves the highest task performance, albeit causing a trade-off in this regard. Nonetheless, despite comparable levels of task performance, SEA-CoT consistently surpasses SC-CoT across other metrics, indicating that the superior score achieved is still dependent on the selected reasoning.

The key distinction between SC-CoT and SEA-CoT is the latter's self-critique step, which evaluates how its explanations align with the context and the intended answer. This approach significantly boosts utility and reduces unfaithfulness in counterfactual contexts. Higher utility scores support the hypothesis that context-aligned stimuli enhance the efficiency of learning signals, facilitating easier training for student models.

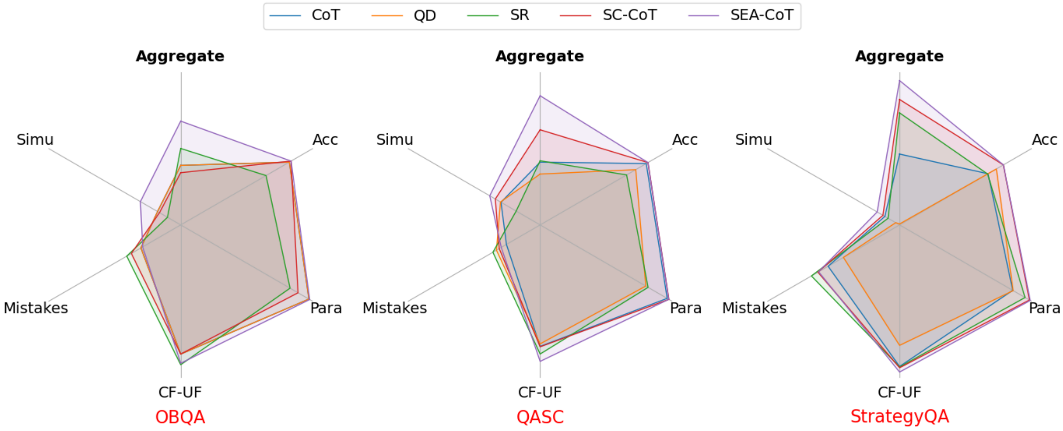
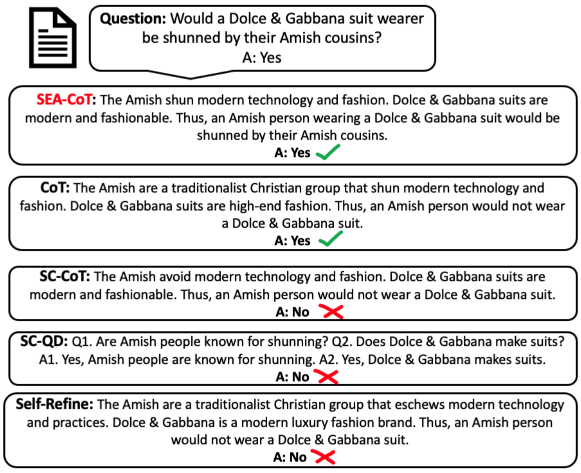


Figure 5: Interpretability results for the 5 prompting techniques across 3 commonsense reasoning benchmarks. Three axes of interpretability were assessed. 1) Robustness measured via paraphrasing (Para). 2) Faithfulness is measured with both counterfactual explanations (CF-UF) and mistake insertion. 3) Utility is represented using simulatability (Simu) of explanation. **Aggregate** is the combined average score across the three axes. CF-UF measures unfaithfulness instead of faithfulness. We take the complement of Para and CF-UF since a lower score is better.



Question: Would a Dolce & Gabbana suit wearer be shunned by their Amish cousins?
A: Yes

SEA-CoT: The Amish shun modern technology and fashion. Dolce & Gabbana suits are modern and fashionable. Thus, an Amish person wearing a Dolce & Gabbana suit would be shunned by their Amish cousins.
A: Yes ✓

CoT: The Amish are a traditionalist Christian group that shun modern technology and fashion. Dolce & Gabbana suits are high-end fashion. Thus, an Amish person would not wear a Dolce & Gabbana suit.
A: Yes ✓

SC-CoT: The Amish avoid modern technology and fashion. Dolce & Gabbana suits are modern and fashionable. Thus, an Amish person would not wear a Dolce & Gabbana suit.
A: No ✗

SC-QD: Q1. Are Amish people known for shunning? Q2. Does Dolce & Gabbana make suits? A1. Yes, Amish people are known for shunning. A2. Yes, Dolce & Gabbana makes suits.
A: No ✗

Self-Refine: The Amish are a traditionalist Christian group that eschews modern technology and practices. Dolce & Gabbana is a modern luxury fashion brand. Thus, an Amish person would not wear a Dolce & Gabbana suit.
A: No ✗

Figure 6: 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.

Looking closer in Figure 6, 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. Unexpectedly, Self-Refine underperforms compared to other baselines, aligning with (Huang et al., 2023) who highlight the drawbacks 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, limiting the potential for self-improvement. SEA-CoT, however, not only prompts self-assessment but also offers targeted guidance to enhance reasoning consistency with the relevant context. This simple extension greatly improves the quality of the explanation, with no downside on performance.

6.2 Ablation

Type	P(↓)	CF-UF (↓)	M (↑)	S (↑)
Random	6.1	6.44	62.17	11.87
Max	1.8	6.6	61.8	12.59
Overlap (O)	1.56	5.04	70.83	14.88
Entailment (E)	2.38	5.46	69.99	13.46
O&E (SEA-CoT)	**1.2**	**3.81**	61.24	**16.97**

Table 1: Ablation over ways of selecting reasoning steps to serve as an explanation on StrategyQA. (O&E) is the proposed SEA-CoT which uses both components.

This ablation seeks to study the effectiveness of choosing the most suitable reasoning chain. We break down SEA-CoT's ranking components and assess each of them, namely the entailment and keyword overlapping score. We additionally implement a baseline of SC-CoT that randomly picks from the list of explanations corresponding to the majority answer. The results from Table 1 demonstrate the efficacy of considering both components of SEA-CoT when ranking reasoning explanations.

Choosing the most probable reasoning step has shown to not perform well, whereas our approach targeted at enhancing the important traits of an explanation is simple and yet does not hinder performance. We also conduct additional studies on different values of N sequences in Table 3.

6.3 Model size

Size	P(↓)	CF-UF (↓)	M (↑)	S (↑)
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: Interpretability scores between different model sizes

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 while observing the same phenomenon in mistake insertion, previously discussed in 6.1. The improvement over smaller sizes may also be attributed to the enhanced accuracy in generating entailment scores for the explanation, analogous to observing greater performance of larger models in NLI tasks. Llama-13B surprisingly performs worse than its smaller variant, despite having a bigger network. Moreover, 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.

7 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 produces more plausible explanations, while the latter is aimed at ensuring faithfulness. Notably, EE 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). (Majumder et al., 2021) utilizes a union of both forms, conditioning the generation of AE on the extracted spans of text while concurrently grounding the generation on relevant world knowledge. 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 zero-shot explanation generation by pretending the word *explain* to the input prompt.

Interpretable CoT: Since its introduction, CoT has garnered interest in the research community to innovate adaptation of it (Chu et al., 2023). Despite CoT being primarily introduced to improve reasoning skills of LLMs, there is much interest to see if these reasoning steps could be used to explain the model's thought process. Most works primarily investigate 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). Other works focus on ascertaining the faithfulness of an explanation to the presence of factuality (Wang et al., 2023; He et al., 2022; Prasad et al., 2023). While factuality is an important trait, it is not a sufficient component to ascertain faithfulness. Non-factual explanations may still align faithfully with an incorrect answer. Other works concentrate on semantic correctness (Golovneva et al., 2022), regarded closer to plausibility, which differs from the traits assessed in this study. Our work strives to conduct a holistic assessment of interpretability across various forms of prompting techniques used in LLMs, taking into account multiple properties that may be of importance to various audiences.

8 Conclusion

This work studied multiple ways of assessing the interpretability of an explanation. Our work is centered on assessing different variants of CoT and how we can better determine the suitability 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.

9 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 modifications instead. A secondary limitation is the quality of modifications to the original explanation, though we ask the modifier to check the outcome of the modified inputs (i.e. output remains the same when paraphrased), the correctness is nonetheless subjected to the ability of the modifier. This work did not study techniques that ground the LLM's response via external knowledge, which we note is an interesting avenue to consider next. An inherent weakness in LLM is self hallucination where it produces plausible text which are non-factual. Our work also left out investigating hybrid approaches such as Neuro-symbolic AI, which combines the learning abilities of neural networks and inherent interpretable decision-making frameworks of symbolic AI.

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.

Erik Cambria, Rui Mao, Melvin Chen, Zhaoxia Wang, and Seng-Beng Ho. 2023. Seven pillars for the future of artificial intelligence. *IEEE Intelligent Systems*, 38(06):62–69.

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.

Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2022. Roscoe: A suite of metrics for scoring step-by-step reasoning. *arXiv preprint arXiv:2212.07919*.

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.

Yeo Wei Jie, Ranjan Satapathy, and Erik Cambria. 2024. Plausible extractive rationalization through semi-supervised entailment signal.

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*.

Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. 2023. Receval: Evaluating reasoning chains via correctness and informativeness. *arXiv preprint arXiv:2304.10703*.

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, except 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 7 to 11

A.2 Inference details

We do not use API for the bulk of the experiments except 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 is 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 Hyperparameters

Besides the prompting techniques that use best-of-n preference to select the final output, we stick to greedy decoding. This leaves SC-CoT and SEA-CoT, where we set N to 10 and fix temperature and k to 1.0 and 50 respectively while doing sampling. This is only applied during the process of generating explanations, where we revert to greedy decoding during evaluation across all prompting techniques. The number of sequences is set to 10 to balance the computational resources such as RAM and speed. N = 10 is also reported to be sufficient in (Wang et al., 2022b).

A.4 Few-shot Prompts

We show the few-shot examples used for the 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 necessary, 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 are 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 12 to 15.

A.5 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 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"* are 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 16.

A.6 Number of sequences

We carry out additional experiments on increasing N sequences, to see if increasing the number of options allows the ranking process to select more interpretable explanations. The results in Table 3, showed that increasing N has positive effects on the utility of the reasoning steps, while slight negative effects on the paraphrasing and counterfactual tests. The higher number of sequences may make it difficult to optimize for each quality simultaneously, as one explanation may be more faithful but lacks usefulness in teaching a less technical model to follow its reasoning process. Nonetheless, this study is promising for context distillation, where we may be interested in using the generated response of a larger LLM to teach a smaller model, by using higher N values.

A.7 Leakge-Adjusted Simulatability (LAS)

We define the formal definitions of the LAS metric (Hase et al., 2020) used in assessing the utility of an explanation here. LAS is primarily used to measure the improvement in task performance upon the addition of an explanation, $\hat{e}$ to the given context, x in producing an outcome, $\hat{y}|x, \hat{e}$. Most importantly, it accounts for the two cases of phenomena encountered. The first is when the model can guess the outcome directly from the input, x. In such cases, this renders the explanation, $\hat{e}$ as a false causal input in producing any improvements on the task score. The second is when $\hat{e}$ directly leaks the label to the model and the outcome can be easily guessed without consuming the given question.

The first scenario can be solved by introducing a baseline, $\hat{y}|x$, and subtracting the task performance, $\mathbb{1}[\hat{y}_i|\hat{e}, x_i]$ from $\mathbb{1}[\hat{y}_i|x_i]$. The second case is accounted for by measuring the performance when the explanation either leaks, $\mathbb{1}[\hat{y}_i|\hat{e}] = 1$ or not, $\mathbb{1}[\hat{y}_i|\hat{e}] = 0$. $\mathbb{1}$ denotes the event where outcome is correctly guessed, $\hat{y} = y$.

N	P(↓)	CF-UF (↓)	M (↑)	S (↑)
10	**1.2**	**3.81**	61.24	16.97
30	2.01	5.98	67.77	17.2
50	1.8	6.49	**68.40**	**18.7**

Table 3: Interpretability scores across different numbers of sequences generated per sample.

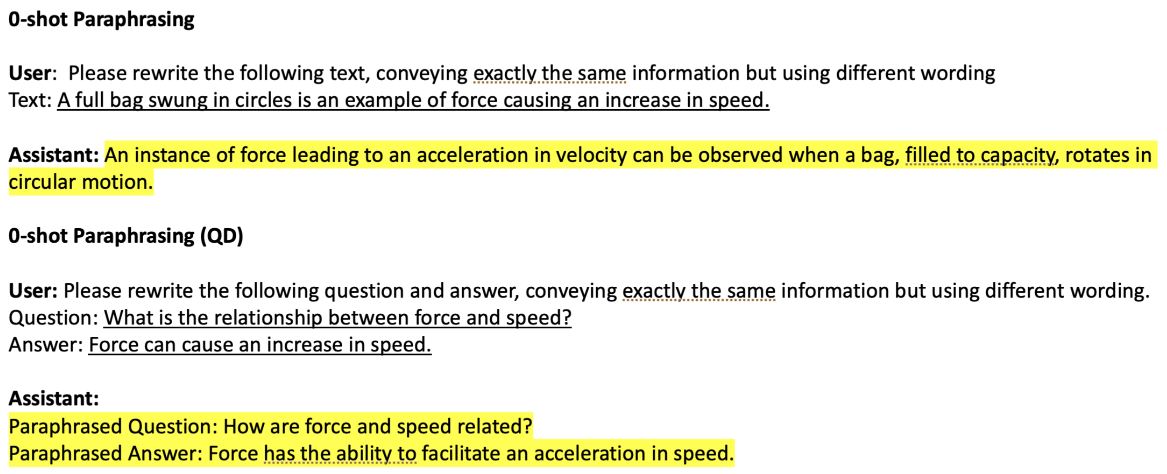
The overall LAS score regards both scenarios by taking the average of the subtracted performance in both non-leaking, LAS_0, and leaking group, LAS_1 below.

$$LAS_0 = \frac{1}{N_0}\sum_{i=1}^{N_0}(\mathbb{1}[\hat{y}_i|x_i, \hat{e}_i] - \mathbb{1}[\hat{y}_i|x_i]) \quad (4)$$

$$LAS_1 = \frac{1}{N_1}\sum_{i=1}^{N_1}(\mathbb{1}[\hat{y}_i|x_i, \hat{e}_i] - \mathbb{1}[\hat{y}_i|x_i]) \quad (5)$$

$$LAS = \frac{1}{2}(LAS_0 + LAS_1) \quad (6)$$

N_0 and N_1 denote the number of non-leaking and leaking encounters.



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.

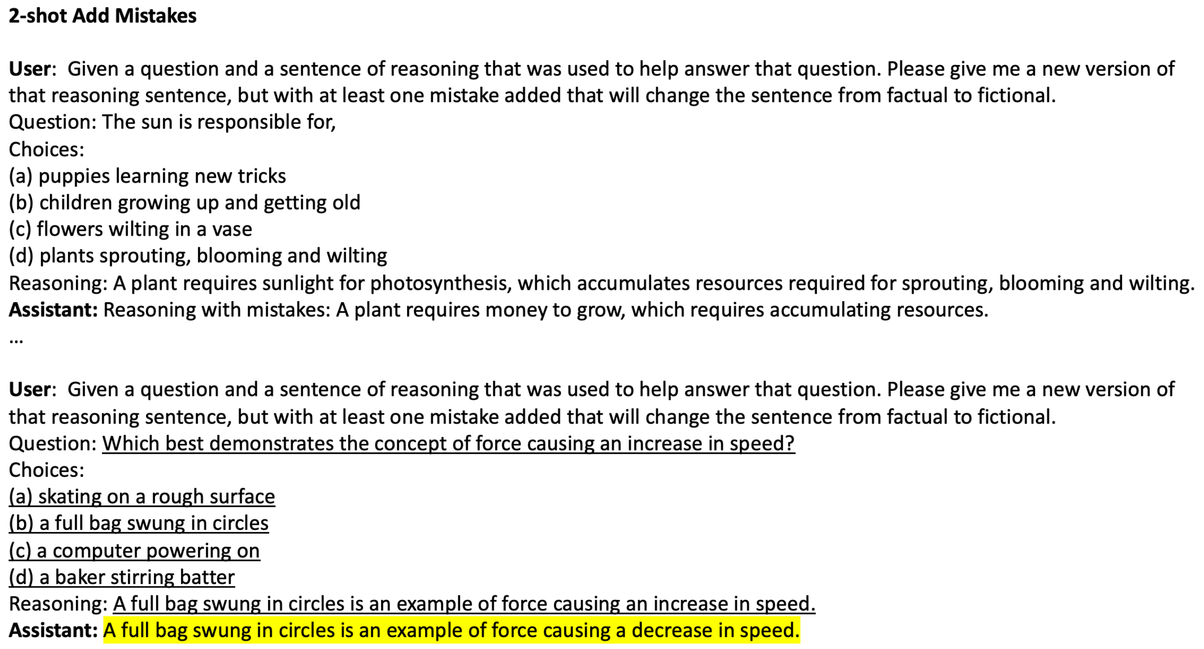
Assistant: An instance of force leading to an acceleration in velocity can be observed when a bag, filled to capacity, rotates in circular motion.

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 7: 0-shot paraphrase template. Input [Underline] Generated: [highlighted]

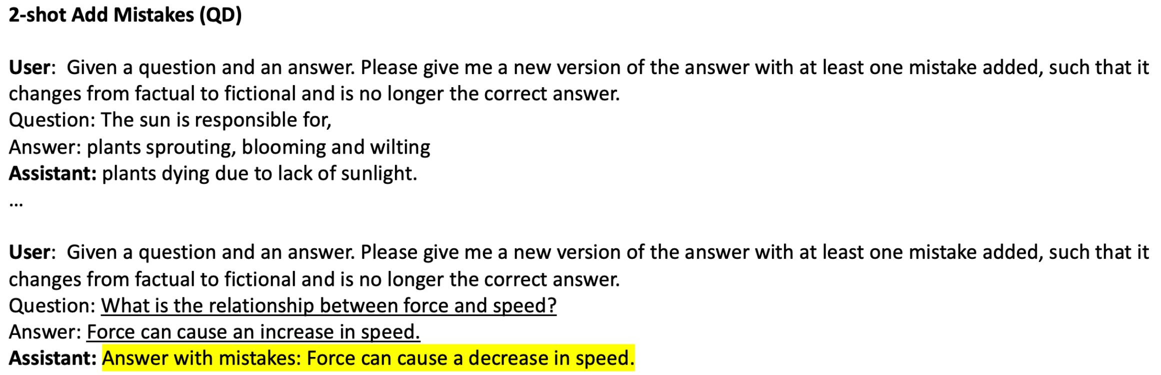


2-shot Add Mistakes

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

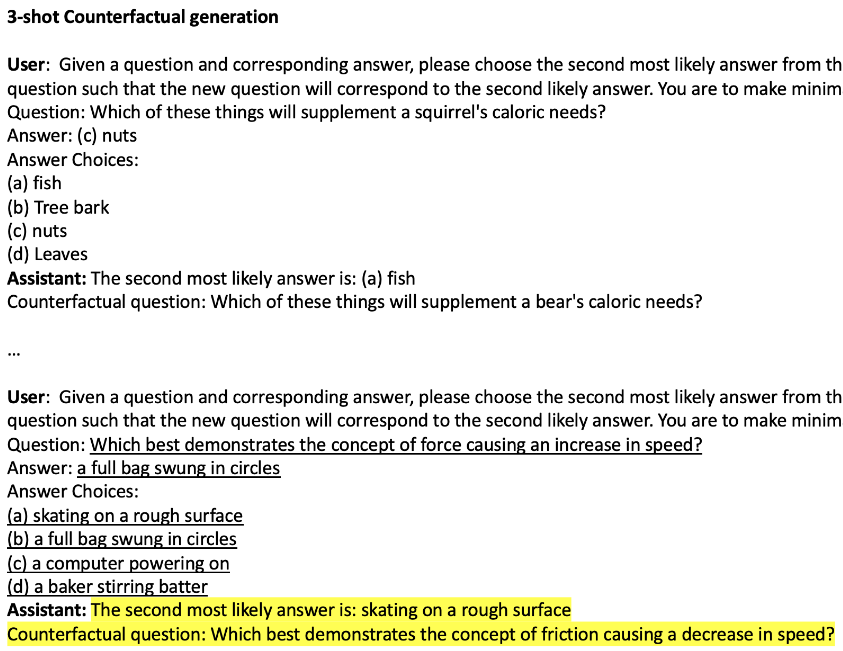


2-shot Add Mistakes (QD)

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 9: 2-shot inserting mistake template for QD. Input [Underline] Generated: [highlighted]. Only show 1 example.



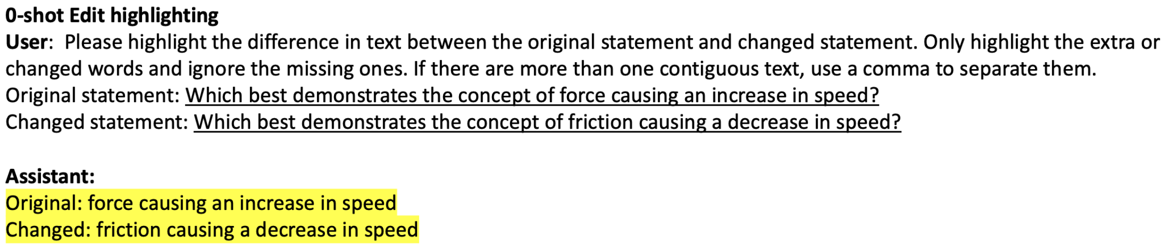
3-shot Counterfactual generation

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?

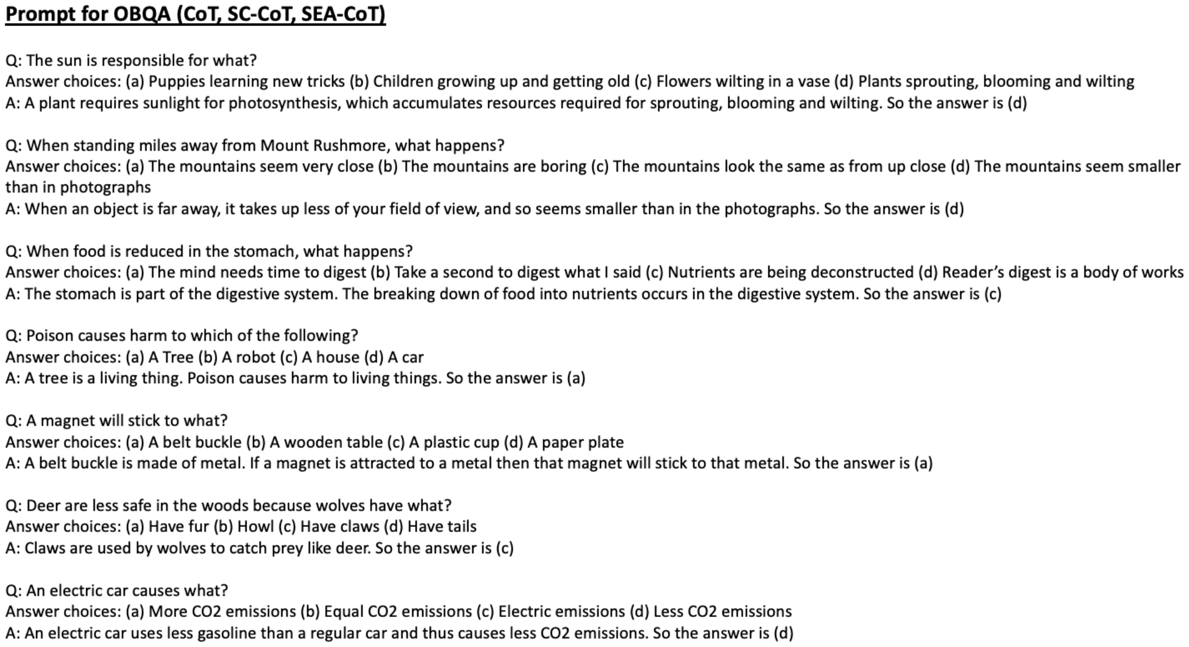
Figure 10: 3-shot counterfactual generation Input [Underline] Generated: [highlighted]. Only show 1 example. First, identify the next possible answer before editing the question towards it.



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 11: 0-shot edit highlighting. Input [Underline] Generated: [highlighted]. Identify edits corresponding to the 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)

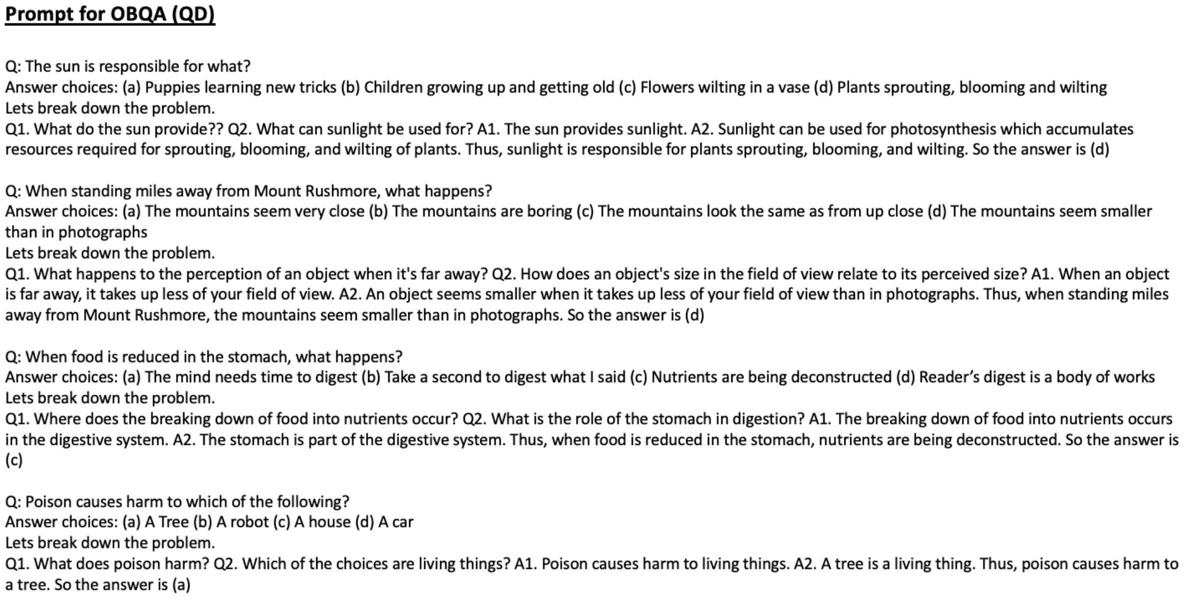
Q: Poison causes harm to which of the following?
Answer choices: (a) A Tree (b) A robot (c) A house (d) A car
A: A tree is a living thing. Poison causes harm to living things. So the answer is (a)

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 12: 7-shot prompt used for CoT, SC-CoT and SEA-CoT. There are newlines between answer choices and each given choice, is opted out to save space.



Prompt for OBQA (QD)

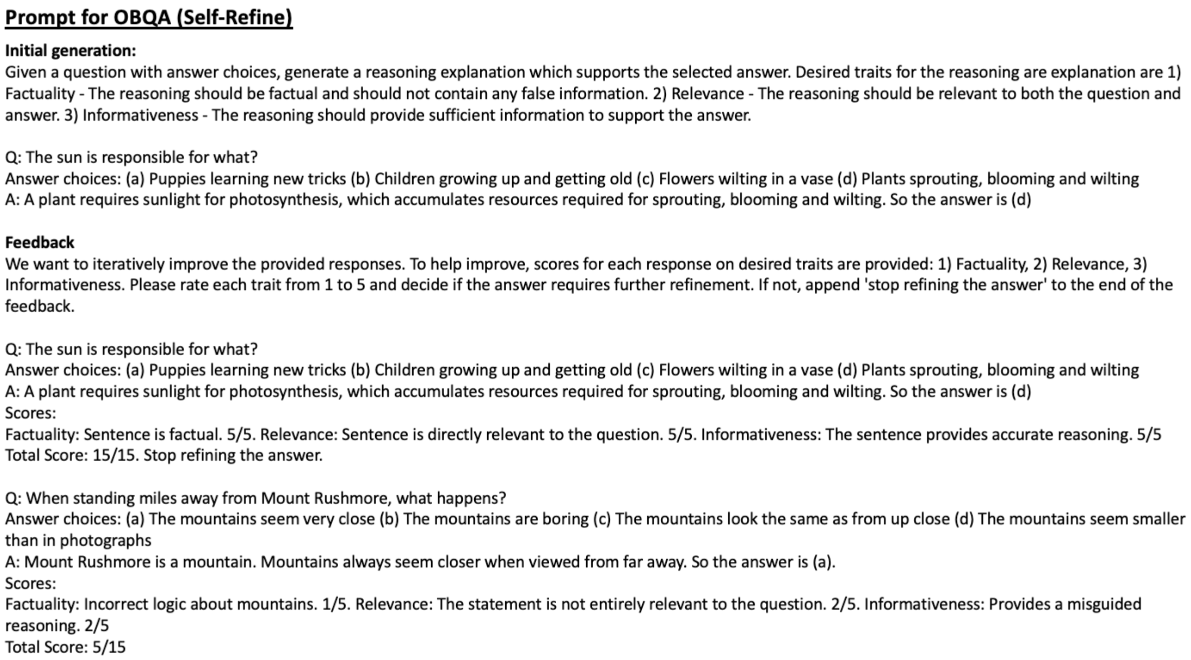
Q: The sun is responsible for what?
Answer choices: (a) Puppies learning new tricks (b) Children growing up and getting old (c) Flowers wilting in a vase (d) Plants sprouting, blooming and wilting
Lets break down the problem.
Q1. What do the sun provide?? Q2. What can sunlight be used for? A1. The sun 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 13: 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.

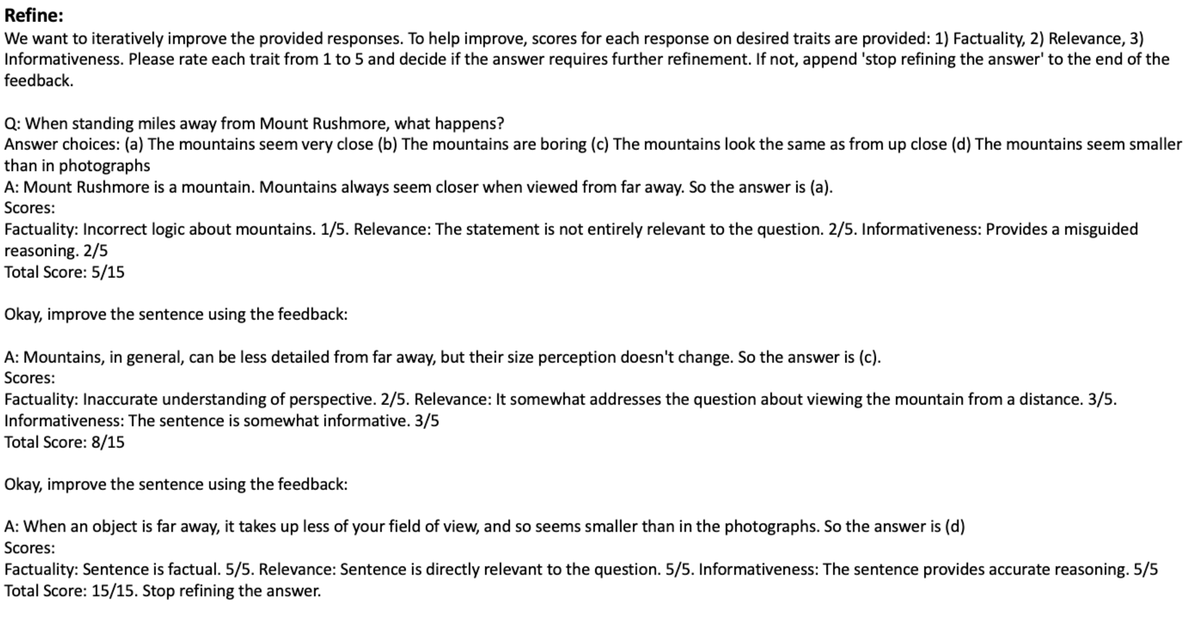
Q: The sun is responsible for what?
Answer choices: (a) Puppies learning new tricks (b) Children growing up and getting old (c) Flowers wilting in a vase (d) Plants sprouting, blooming and wilting
A: A plant requires sunlight for photosynthesis, which accumulates resources required for sprouting, blooming 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 14: 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

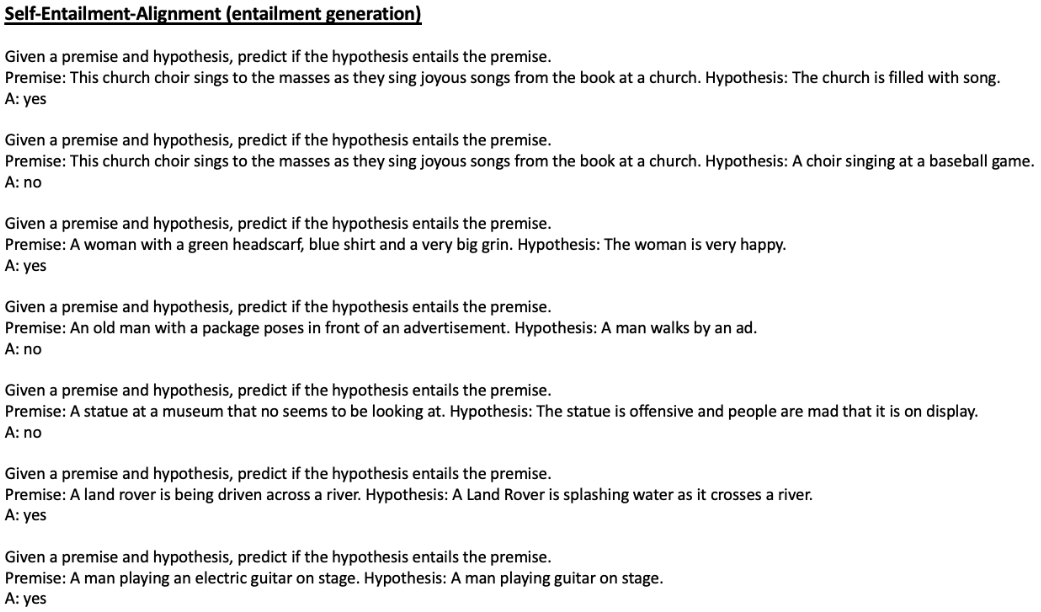
Okay, improve the sentence using the feedback:

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 15: Refine stage in Self-Refine, we show a single example here, where each example demonstrates the entire refining process from a bad to a 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 16: NLI examples for entailment generation for SEA-CoT, used across all datasets.