It’s Not Easy Being Wrong: Large Language Models Struggle with Process of Elimination Reasoning

Nishant Balepur  Shramay Palta  Rachel Rudinger
University of Maryland, College Park, USA
{nbalepur, spalta, rudinger}@umd.edu

Abstract

Chain-of-thought (COT) prompting can help large language models (LLMs) reason toward correct answers, but its efficacy in reasoning toward incorrect answers is unexplored. This process of elimination (PoE), when used with COT, can enhance self-consistency, interpretability, and tasks such as medical diagnoses of exclusion. Thus, we propose PoE with COT, where LLMs must reason toward incorrect options on multiple-choice questions. We evaluate the ability of GPT-3.5, LLaMA-2, and Falcon to perform PoE with COT on a total of four commonsense and scientific reasoning datasets. We find that the strategy of PoE always underperforms the strategy of choosing the correct answer. The agreement of these strategies is also lower than the self-consistency of each strategy. To study these issues further, we conduct error analyses and give suggestions for future work.

1 Introduction

Recent research has aimed to unlock the reasoning capabilities of large language models (LLMs) (Nye et al., 2022). As part of this effort, researchers have proposed techniques such as chain-of-thought (COT) prompting to help LLMs verbally reason toward correct answers (Wei et al., 2022b; Kojima et al., 2022). Such reasoning can improve the accuracy and interpretability of LLM decision-making (Creswell et al., 2023; Huang and Chang, 2023).

While several works use COT to select correct answers (Figure 1 blue), they do not study if COT can identify incorrect answers (Figure 1 red). Thus, drawing from test-taking strategies for multiple-choice (MC) exams (Tversky, 1972), we propose a new task to probe LLM reasoning: process of elimination (PoE) with COT. Previous works have performed PoE by discarding low-confidence options when selecting correct answers (Ma and Du, 2023), but whether generative LLMs can directly reason toward incorrect options with COT is unexplored.

Analyzing PoE with COT poses several benefits. First, it can give insights into the consistency of LLM reasoning (Wang et al., 2023). For 2-choice MC questions, we can study if the decisions from PoE with COT and directly answering the question with COT agree, as both should arrive at the same answer (§4.2). Second, PoE with COT can enhance LLM interpretability. COT reasoning is often seen as a rationale as to why the model thinks an option is correct, but users may also want to know why the model thinks alternatives are incorrect. Third, many applications benefit from PoE, such as reaching medical diagnoses via exclusion (Fred, 2013; Kline, 2018), troubleshooting by eliminating root causes (Gutoff and Cohen, 2006; Gugerty, 2007), and ruling out scientific hypotheses using new evidence (Norton, 1995; Forber, 2011). Thus, to enable interpretable LLMs in domains like medicine, customer service, and research, we must know: Do LLMs have the ability to perform PoE with COT?

To study this question, we independently prompt the GPT-3.5 (Brown et al., 2020), LLaMA-2 (Touvron et al., 2023), and Falcon (Penedo et al., 2023) LLMs to directly select the correct option and eliminate the incorrect one, with and without COT, on
We introduce PoE with COT, a new reasoning strategy with COT to suggest directions for future research. We study selecting correct and incorrect options in 2-choice commonsense (Sap et al., 2019; Talmor et al., 2019) and scientific (Clark et al., 2018; Michaylov et al., 2018) reasoning datasets. In this 2-choice setup, we can juxtapose LLM reasoning abilities in direct answer selection against PoE, as ideal LLMs would always have both strategies agree.

We find that PoE always underperforms direct answer selection for our LLMs, but this gap narrows as model size scales, implying that PoE with COT may only be attainable for larger LLMs (§4.1). Further, these two strategies do not reliably arrive at the same answer, revealing an inherent inconsistency in LLM reasoning (§4.2). For more insights, we analyze the errors in the rationales of PoE with COT, finding that most stem from reasoning errors and task misalignment (§4.3), and show how these 2-choice errors propagate when using PoE on full MC questions (§4.4). Based on our results, we give suggestions for future work. Our contributions are:

1) We introduce PoE with COT, a new reasoning task for LLMs that can benefit interpretability, self-consistency, and downstream applications.
2) We benchmark the abilities of GPT, LLaMA-2, and Falcon to perform PoE with and without COT on scientific and commonsense reasoning datasets.
3) We evaluate the consistency and errors of PoE with COT to suggest directions for future research.

2 Problem Definition

We study selecting correct and incorrect options in MC settings, where we are given a question $q$ and $n$ choices $C$, exactly one of which is correct. To simplify our discussion, we define two strategies:

1) Direct Answer (DA) Strategy: The LLM aims to select the correct answer choice $c_f \in C$.

2) Process of Elimination (PoE) Strategy: The LLM aims to select an incorrect choice $c_f \in C$.

We ideally want to perform PoE iteratively, generalizing the strategy for any number of choices. But as a prerequisite, we must know: Are LLMs inherently capable of reasoning toward incorrect choices? To answer this question, we analyze the logical consistency of LLMs via 2-choice questions ($C = \{c_A, c_B\}$), as a robust LLM would always have DA and PoE agree: predicting that one choice is correct entails that the other is incorrect. If PoE fails with two choices, it reveals that LLMs are inherently weaker at picking incorrect options and these errors will propagate when using PoE iteratively. We first show that PoE is unreliable in 2-choice settings (§4.1), and later show this leads to error propagation in iterative settings (§4.4).

3 Experimental Setup

3.1 Strategy Implementation

To provide sufficient context for the LLMs, we use few-shot (10) prompts to implement the strategies, leaving the analysis of 0-shot PoE with COT for future work. We construct the following prompts:

1) DA Base and PoE Base: These prompts task the LLM with performing the DA and PoE strategies without reasoning. The in-context examples in these prompts follow a similar format to Figure 1, but only the correct/incorrect answer choice ((A) or (B)) follow the “Correct/Incorrect Answer:” labels.

2) DA COT and PoE COT: These prompts add COT to the DA and PoE Base prompts. The few-shot examples follow the same format as Figure 1, requiring the LLM to give a step-by-step rationale before answering. Notably, the DA and PoE COT examples are created to be distinct such that DA/PoE reason toward the correct/incorrect answer without discussing the other choice’s validity. This is meant to prevent unhelpful PoE rationales like “The answer is (A). So the incorrect answer is (B).”

To study if COT reasoning improves the accuracy of each strategy (§4.1), we can compare the accuracy of the Base prompts with their COT counterparts (e.g. PoE Base vs PoE COT). Further, comparing the outcomes of the DA and PoE prompts (e.g. DA COT vs PoE COT) allows us to measure the LLMs’ logical consistency in choosing correct answers and eliminating incorrect options (§4.2).

The COT rationales are written by one Ph.D. student and verified by a second (the authors). We also prepend a natural language instruction to the prompt explaining the strategy. More prompting details and examples can be found in Appendix A.2.

3.2 Datasets

We examine multiple-choice commonsense and scientific reasoning questions, given that the DA strategy has shown to perform better on these questions using COT reasoning (Wei et al., 2022b; Lu et al., 2022a; Zhang et al., 2023b; Lanham et al., 2023). Commonsense reasoning questions have models reason about knowledge understood by most humans. We use Social IQa (Sap et al., 2019, SIQA),

---

2 We do not use binary Yes/No datasets, like BoolQA (Clark et al., 2019), as the reason for why a correct answer is correct and an incorrect answer is incorrect are the exact same. Thus, they do not align with our requirement of distinct rationales.

10144
containing questions about social interactions, and CommonsenseQA (Talmor et al., 2019, CQA), concerning knowledge encoded in ConceptNet.

Scientific reasoning datasets have models make logical inferences based on scientific facts. We use ARC (Clark et al., 2018) and OpenBookQA (Mihaylov et al., 2018, OBQA), testing factual recall, multi-step reasoning, and some commonsense.

We sample 500 random questions from the test set of each dataset for evaluation, and sample 10 random training examples for the 10-shot prompts. To leave each question with only two choices, we randomly eliminate all but one incorrect choice. Dataset statistics can be found in Appendix A.1.

3.3 Models
We study three families of LLMs. First, we use GPT-3.5 (gpt-3.5-turbo-0613) (Ouyang et al., 2022) with the OpenAI API. We also use two open-source base LLMs: LLaMA-2 (7B, 13B, 70B) (Touvron et al., 2023) and Falcon (7B, 40B, 4-bit 180B) (Penedo et al., 2023), loaded via huggingface. Each LLM decodes with 0.3 temperature.

4 Results
4.1 Can LLMs Perform PoE?
In Figure 2, DA Base and DA COT surpass their PoE counterparts for every model and dataset. Notably, smaller LLMs (7B, 13B) using PoE often underperform a 50/50 guessing model, suggesting that the smaller models are fully misunderstanding the PoE task, despite being given 10 examples. In contrast, the accuracy discrepancy between the PoE and DA strategies is less pronounced for larger LLMs, implying that PoE may only be an attainable ability for larger LLMs (Wei et al., 2022a). Our results show that LLMs, especially smaller models, may have a bias toward choosing correct answers, and thus struggle to perform PoE accurately. We speculate this occurs because LLM pre-training data likely contains many MCQA questions and explanations justifying the correct choice, but far fewer explanations justifying incorrect choices, leading to a bias towards choosing correct answers.

Further, while COT tends to slightly improve the accuracy of the DA strategy across LLMs, this is not true for PoE. Specifically, DA COT has equal or better accuracy than DA Base in 19/24 cases (8/12 for LLaMA, 7/12 for Falcon, 4/4 for GPT). However, while PoE COT surpasses PoE Base in 10/12 cases for LLaMA and 2/4 cases for GPT, this improvement is never seen for Falcon, indicating a weakness in reasoning. This discrepancy across models further motivates PoE as a promising task. While COT tends to improve or maintain accuracy when picking correct answers, its accuracy wavers by LLM when eliminating choices. Thus, we believe that the accuracy of PoE with COT could be used to evaluate the reasoning abilities of LLMs.

4.2 Are PoE and DA Consistent?
In Table 1, we study the logical consistency of DA and PoE, i.e., if the strategies converge to the same answer. For every LLM and dataset, we find that DA more often agrees with itself upon a repeated inference (i.e. self-consistency) than agrees with its PoE counterpart. This suggests that LLMs have an innate logical inconsistency when asked to execute these two strategies, which cannot be ascribed to sampling variation from our temperature selection.

We suggest three future directions based on our...
Table 1: Agreement of DA strategies (Base/COT) with PoE counterparts vs. Self-consistency. More consistent methods are underlined. Full results in Appendix B.2.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>DA Base PoE Self</th>
<th>DA COT PoE Self</th>
<th>DA Base PoE Self</th>
<th>DA COT PoE Self</th>
</tr>
</thead>
<tbody>
<tr>
<td>CQA</td>
<td>91.4 97.6</td>
<td>78.2 92.9</td>
<td>81.3 97.3</td>
<td>86.0 97.6</td>
</tr>
<tr>
<td>SIQA</td>
<td>86.6 95.0</td>
<td>69.4 91.4</td>
<td>80.7 97.0</td>
<td>76.2 95.2</td>
</tr>
<tr>
<td>ARC</td>
<td>89.3 98.5</td>
<td>79.1 90.8</td>
<td>86.6 97.5</td>
<td>85.5 97.0</td>
</tr>
<tr>
<td>OBQA</td>
<td>89.7 96.2</td>
<td>77.2 95.1</td>
<td>72.3 98.2</td>
<td>81.4 96.8</td>
</tr>
</tbody>
</table>

The agreement of DA and PoE can assess logical robustness. Future LLMs can aim to achieve DA/PoE consistency closer to DA self-consistency. Second, measuring when DA and PoE agree could be used for LLM confidence calibration (Cheng et al., 2023), as an LLM that arrives at the same solution with diverse strategies may be more confident. Lastly, as LLMs have shown to improve rationales by combining multiple reasoning chains (Yoran et al., 2023), future works could similarly try to synthesize DA and PoE reasoning chains.

4.3 Why does PoE with COT Fail?

To study the issues of PoE, we examine 50 rationales from CQA and ARC where PoE COT failed. We use our three largest LLMs, deemed the best at PoE. Our error types are: 1) Misaligned Rationale: Justifies a choice as correct instead of incorrect; 2) Reasoning Error: Error in the reasoning chain, such as an inaccurate premise or conclusion; 3) No Breakdown: Restates choice without breakdown into reasoning steps; 4) Hallucination: References non-existent parts of the question; 5) PoE N/A: Question is unsuitable for PoE; and 6) Dataset: Dataset quality issue. Examples of rationales with each error type are in Appendix Table 14.

In Figure 3, most errors stem from reasoning or misaligned rationales. The prevalence of these errors suggests that our PoE COT setting reveals the weaknesses of our tested LLMs to reason and follow in-context instructions, making our task a suitable testbed for these abilities. Notably, many misaligned rationales occur with negated questions (e.g. “What would Grace not do?”). LLMs struggle to reason under negation (Ravichander et al., 2022) and since PoE is a negated reasoning technique, we may expect LLMs to underperform with this “double negation.” One solution to these issues is to fine-tune LLMs on PoE rationales. This may enhance PoE, but it would be interesting to see if this could also bolster overall reasoning capabilities.

4.4 Is Iterative PoE Reliable?

We speculated that if PoE fails with two choices, errors will propagate with full MC questions. To confirm this, we run the setup in §3 on ARC/CQA with LLaMA, eliminating \( i \in [0, n-1] \) choices and then selecting the correct answer. The accuracy of this strategy decreases with each iteration both with PoE Base and COT (Figure 4). Thus, we suggest that future works improve the robustness of LLMs in 2-choice settings before trying iterative PoE, as iterative PoE currently leads to error propagation.

5 Related Work

LLM Reasoning: Several works explore if LLMs can reason with a chain-of-thought (COT) before giving a final answer (Wei et al., 2022b; Huang and Chang, 2023). Recent works on COT probe its faithfulness (Lyu et al., 2023; Lanham et al., 2023) and planning (Saparov and He, 2023). Similar to PoE with COT, prompt techniques like debate (Michael et al., 2023) and maieutic prompting (Jung et al., 2022) produce supporting rationales for multiple options or let the model make its own decision, but we are the first to pinpoint if LLMs can reason why options are incorrect, rather than cor-
rect. Appendix C discusses these works in-depth. **Process of Elimination (PoE):** PoE has been studied in various settings. Parikh et al. (2018) train a neural model to perform PoE for reading comprehension. Zhang et al. (2024) and Ma and Du (2023) use LMs to find the probability that candidates should be excluded for prompt ensembling and MCQA. Tong et al. (2023) design a pipeline where an LLM proposes answers and removes incorrect proposals. While these works discard options to pick correct answers, we assess if generative LLMs can reason toward incorrect options with COT. PoE with COT also has similarities with flipped classification label studies (Wei et al., 2023), but we do not override the semantics of labels.

6 Conclusion

We study COT with process of elimination (PoE) and direct answer (DA) strategies in 2-choice commonsense and scientific reasoning datasets. We find that PoE underperforms DA in accuracy across all tested LLMs, but this gap narrows as model size scales. We also explore the logical inconsistencies between DA and PoE and categories of errors present in PoE rationales. Based on our results, we suggest several research directions: using PoE with COT as LLM reasoning/consistency benchmarks, combining DA and PoE with COT for calibration and refined rationales, and fine-tuning LLMs for PoE to enhance overall reasoning capabilities.

7 Limitations

LLMs are sensitive to prompts (Lu et al., 2022b), so the accuracy of PoE COT could vary depending on the number of in-context examples and the human-written rationales in these prompts (Wei et al., 2022b). To mitigate this, we use an ample number (10) of in-context examples and write our prompts following the same format as Wei et al. (2022b). Each rationale is written by one Ph.D. student and verified by a second Ph.D. student, both of whom work on LLM research, to ensure the rationales are high-quality. Rationales were also edited after running inference and identifying problems on a small validation set. Future works could explore rationale refinement or self-correction techniques (Zhang et al., 2023a; Fu et al., 2023; Huang et al., 2024) to have LLMs create their own prompts.

Further, while PoE should ideally be able to give a rationale as to why one choice is correct without mentioning the other choice, this may be infeasible in certain scenarios. For example, given a mathematics question like “What is 2+2*2?” with the choices “4” and “6”, the chain-of-thought leading the correct answer is clear (e.g. “2 * 2 is 4 and 2 + 4 is 6, so the correct answer is 6”). However, it is very difficult to justify why “4” is incorrect without knowing that “2+2*2=6”. To try to avoid these types of questions, we only look at scientific and commonsense reasoning questions, as we find it to be straightforward to explain why incorrect choices are incorrect. We also use these datasets to test PoE on questions with varying levels of objectivity (i.e. commonsense questions are more subjective than scientific questions). PoE is one of many test-taking strategies, so similar to students, future works could try teaching LLMs to choose the best strategies when answering multiple-choice questions, overcoming this limitation.

Lastly, due to resource constraints, we have not tested the ability of GPT-4 to perform PoE. In future studies related to PoE, it could be beneficial to evaluate GPT-4 to see if the LLM outperforms GPT-3.5 when executing PoE. If the gap between DA and PoE is smaller for GPT-4 than GPT-3.5, it would further support the idea that larger and more capable LLMs are more adept at performing PoE.

8 Ethical Considerations

Process of elimination is a strategy that aims to help LLMs eliminate incorrect options, with the goal of promoting interpretability and self-consistency, as well as enhancing downstream applications. However, our current findings suggest that PoE with COT may be an unreliable technique. Thus, we advise researchers and LLM practitioners to take caution before deploying PoE in any application.

Further, as with all reasoning techniques, researchers should ensure that PoE rationales are not based on stereotypes or biases. This is especially important in some of the downstream applications we mention in this work, like medical diagnoses of exclusion, where the backgrounds of certain individuals could introduce bias into the LLM’s reasoning. We hope that future research endeavors will jointly attempt to address these biases and improve the robustness of LLM reasoning, fostering more equitable decision-making frameworks.

9 Acknowledgements

We would like to thank members of the CLIP lab at the University of Maryland and external collabora-

10147
tors for their feedback and discussions of this work, including Dayeon (Zoey) Ki, Yu (Hope) Hou, Yoo Yeon Sung, Shi Feng, and Jordan Boyd-Graber. We also thank the anonymous reviewers for their feedback. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE 2236417. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References


Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022a. Learn to explain: Multimodal reasoning via thought chains for science


A Experimental Setup Details

A.1 Dataset Statistics
In Table 2, we display summary statistics for the datasets used in our experiments. All datasets are publicly available and free to use on HuggingFace.

A.2 Prompt Details
Examples of the prompts used for DA COT and PoE COT on each dataset can be found in Figures 6, 7, 8, 9, 10, 11, 12, and 13. The Base versions of these prompts follow the same format, except only the answer choice follows the “Correct Answer:” and “Incorrect Answer:” labels. For example, if the question was “Where is Chicago?” with choices “Illinois” and “the Moon,” the DA Base prompt would be:

Question: Where is Chicago?
Choices:
(A) Illinois
(B) The Moon
Correct Answer: (A)

and the PoE Base prompt would be:

Question: Where is Chicago?
Choices:
(A) Illinois
(B) The Moon
Incorrect Answer: (B)

Before the in-context learning examples, we prepend the instruction “Your goal is to identify the correct answer to the multiple choice question” to the DA prompts, and the instruction “Your goal is to identify the incorrect answer to the multiple choice question” to the PoE prompts. For the 10 total in-context examples, five of the examples select choice (A) and five of the examples select choice (B), which are shuffled. This ensures that our few-shot prompts are as strong as possible. We release the full prompts along with our code.

A.3 Model Implementation
The 7B open-source LLMs were run on a single NVIDIA RTX A6000 GPU. The other open-source LLMs were run on 8 NVIDIA RTX A5000 GPUs. GPT-3.5 was run using CPU only. We allocated 24 hours for a single LLM to run all four strategies (DA Base, DA COT, PoE Base, PoE COT) on all four datasets. The experiments from Figure 2 are from a single run, while the experiments from Table 1 require up to two runs.

B Detailed Results

B.1 Detailed Quantitative Evaluation
In Tables 3, 4, 5, and 6, we provide the numerical results from Figure 2. We also calculate the difference in accuracies of the DA and POE strategies, along with their associated p-values. We find that a majority of the differences in accuracy between DA and PoE are statistically significant.

B.2 Full Consistency Experiments
We report the consistency results from §4.2 for all LLMs and datasets. In addition to the self-consistency of DA, we include the self-consistency of PoE to ensure that the lack of consistency between DA and PoE is not due to higher sampling variance for PoE. In Tables 7, 8, and 9, we report the results for LLaMA-2. In Tables 10, 11, and 12, we report the results for Falcon. In Table 13, we report the results for GPT-3.5.

We find that the agreement of DA and PoE is the lowest across all datasets and LLMs, except for LLaMA-2 70B on Commonsense QA. We also note that PoE self-consistency is typically lower than DA self-consistency, meaning that PoE has a higher variance which further points to its unreliability.

B.3 Qualitative Results
Examples of PoE rationales that fall into the categories of our error analysis can be found in Table 14. Examples of sound rationales generated by PoE COT can be found in Table 15. In Figure 5, we compare the distribution of errors of PoE COT and DA COT. For DA COT, we use the same error categories defined in §4.3, except “PoE N/A” is replaced with “DA N/A”. Further, we only evaluate 20 DA COT rationales, since there were fewer instances of errors from DA COT to choose from.

The majority of the DA COT errors were reasoning errors (and not misaligned rationales), suggesting that the LLMs better understand the DA task compared to PoE. This may further indicate that LLMs have a training bias toward identifying correct answers. Further, we find that many of the misaligned rationale errors that occur with DA COT are also due to negated questions (e.g. “What would Grace not do?”), reinforcing the difficulty of LLMs to properly address questions with negation.
B.4 Iterative PoE Details

In this section, we provide more details on our setup for the iterative PoE experiment in §4.4. If a question has $n$ choices, we independently prompt the LLM using PoE Base and PoE COT to eliminate an answer $n-1$ times, so the model is unaware of its previous decisions. At each step when there are $i \in [2, n]$ choices, we also run the DA Base and DA COT prompts, corresponding to every possible use case of PoE (i.e. eliminate $i$ choices and then choose the correct answer). PoE Base only is used with DA Base, and PoE COT is only used with DA COT.

In Tables 16 and 17, we display the results for iterative PoE on Commonsense QA and ARC, respectively. Overall, we find that the raw accuracy decreases significantly as more choices get eliminated, confirming our intuition that error propagation is an issue in this setting. Although PoE has higher accuracy than DA when the number of choices remaining is higher, this finding is not that significant when you consider the probability of performing each strategy accurately by chance. For example, when there are 5 choices, there is a probability of $\frac{4}{5} = 0.8$ of guessing an incorrect answer correctly, while there is only a probability of $\frac{1}{5} = 0.2$ of guessing the correct answer correctly. If future research eventually leads to LLMs that can perform the DA and PoE strategies at similar abilities, it would be interesting to revisit the reliability and effectiveness of iterative PoE.

B.5 Why not compare with score-based PoE techniques?

As described in the related work, previous works have explored PoE as a means to discard low-confidence options when selecting correct answers (Zhang et al., 2024; Ma and Du, 2023). These works execute PoE by first prompting an LM to choose the correct answer, and obtain a distribution of token probabilities over all of the options. The answers with token probabilities below a certain threshold are discarded. For example, Ma and Du (2023) discard answers with probabilities that are lower than the average probability of all answers.

This approach, however, is not very useful in the two-choice setting that we study, as the model will always be consistent. To illustrate, in a 2-choice setting, the LM will ascribe choice “A” probability $a$ and choice “B” probability $b$ (where $a + b = 1$). Assuming $a > b$ without loss of generality, a direct answer strategy in this context would always select choice “A” (since $a > b$), and PoE would always eliminate choice “B” (since $b < \frac{1}{2}(a + b)$).

Thus, comparing with these score-based PoE techniques does not give us any useful signals into the reasoning and decision-making capabilities of LLMs, as the model will always be logically consistent. In the PoE and DA setup that we describe in this work, models have the potential to be logically inconsistent, providing deeper insights into the true decision-making capabilities of our tested LLMs. Further, the aspect of COT allows us to interpret the rationales from models (§4.3), which cannot be accomplished with score-based PoE techniques.

C Related Work

There exist a set of techniques, such as Maieutic Prompting (MP) (Jung et al., 2022), that also employ LLMs to generate rationales for multiple options in multiple-choice question answering. Below, we describe the novelty of PoE with COT and how our task is different from the setup of MP.

In MP, the LLM is asked to generate unique rationales of why a True/False question (e.g. War cannot have a tie?) is true and why it is false. In other words, MP forces the model to generate supporting rationales for both a correct answer and an incorrect answer. This is subtly but critically distinct from our setup. In PoE with CoT, the model is given a (non-True/False) question with two answer choices, and is asked to select the incorrect choice and provide a rationale for why it is incorrect. In short, MP forces a model to explain why an incorrect (T/F) answer is correct, while we ask the model to select an incorrect (non-T/F) answer and explain why it is incorrect.

For example, given the T/F question “Is 1 greater than 2,” MP would ask a model to explain (impossibly) why it is true that 1 is greater than 2. Conversely, given the non-T/F question “Where do you put your grapes before checking out,” PoE would ask a model to reason why you don’t put grapes in your mouth before paying at a grocery store (per Figure 1). These setups are inherently different.

Other techniques, such as debate (Michael et al., 2023), also fall into this same category as MP. Further, in debate, is made even more explicit in debate that LLMs may be unreliable and inaccurately arguing for an incorrect option. In contrast, PoE with COT tasks LLMs with generating accurate rationales as to why an answer is incorrect.
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Category</th>
<th># Questions</th>
<th># Choices</th>
<th>Proportion where Gold Answer is Choice (A)</th>
<th>Avg Question Length</th>
<th>Avg Choice Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Commonsense QA</td>
<td>Commonsense Reasoning</td>
<td>500</td>
<td>2</td>
<td>0.514</td>
<td>15.60</td>
<td>2.06</td>
</tr>
<tr>
<td>Social IQa</td>
<td>Commonsense Reasoning</td>
<td>500</td>
<td>2</td>
<td>0.490</td>
<td>24.21</td>
<td>4.09</td>
</tr>
<tr>
<td>ARC</td>
<td>Scientific Reasoning</td>
<td>500</td>
<td>2</td>
<td>0.486</td>
<td>25.78</td>
<td>6.31</td>
</tr>
<tr>
<td>OpenBook QA</td>
<td>Scientific Reasoning</td>
<td>500</td>
<td>2</td>
<td>0.495</td>
<td>11.72</td>
<td>3.72</td>
</tr>
</tbody>
</table>

Table 2: Summary statistics for the datasets used in our experiments. Average length is calculated with the GPT-4 tokenizer, implemented through the tiktoken library.

<table>
<thead>
<tr>
<th>Model</th>
<th>DA Base</th>
<th>PoE Base</th>
<th>Base Diff</th>
<th>Base p-val</th>
<th>DA COT</th>
<th>PoE COT</th>
<th>COT Diff</th>
<th>COT p-val</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLaMA 7B</td>
<td>0.816</td>
<td>0.320</td>
<td>0.496</td>
<td>0.000</td>
<td>0.862</td>
<td>0.458</td>
<td>0.404</td>
<td>0.000</td>
</tr>
<tr>
<td>LLaMA 13B</td>
<td>0.858</td>
<td>0.266</td>
<td>0.592</td>
<td>0.000</td>
<td>0.868</td>
<td>0.421</td>
<td>0.447</td>
<td>0.000</td>
</tr>
<tr>
<td>LLaMA 70B</td>
<td>0.930</td>
<td>0.846</td>
<td>0.084</td>
<td>0.000</td>
<td>0.916</td>
<td>0.808</td>
<td>0.108</td>
<td>0.000</td>
</tr>
<tr>
<td>Falcon 7B</td>
<td>0.516</td>
<td>0.472</td>
<td>0.044</td>
<td>0.164</td>
<td>0.690</td>
<td>0.466</td>
<td>0.224</td>
<td>0.000</td>
</tr>
<tr>
<td>Falcon 40B</td>
<td>0.892</td>
<td>0.740</td>
<td>0.152</td>
<td>0.000</td>
<td>0.876</td>
<td>0.532</td>
<td>0.344</td>
<td>0.000</td>
</tr>
<tr>
<td>Falcon 180B</td>
<td>0.924</td>
<td>0.912</td>
<td>0.012</td>
<td>0.490</td>
<td>0.926</td>
<td>0.786</td>
<td>0.140</td>
<td>0.000</td>
</tr>
<tr>
<td>GPT 3.5</td>
<td>0.890</td>
<td>0.778</td>
<td>0.112</td>
<td>0.000</td>
<td>0.937</td>
<td>0.853</td>
<td>0.084</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Table 3: Evaluation of Direct Answer (DA) and Process of Elimination (PoE) strategies on Commonsense QA, with and without chain-of-thought reasoning. Diff denotes the difference between the accuracies of the two strategies. p-val corresponds to a difference in means t-test. Underlined values denote p-val ≤ 0.05

<table>
<thead>
<tr>
<th>Model</th>
<th>DA Base</th>
<th>PoE Base</th>
<th>Base Diff</th>
<th>Base p-val</th>
<th>DA COT</th>
<th>PoE COT</th>
<th>COT Diff</th>
<th>COT p-val</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLaMA 7B</td>
<td>0.762</td>
<td>0.440</td>
<td>0.322</td>
<td>0.000</td>
<td>0.758</td>
<td>0.548</td>
<td>0.210</td>
<td>0.000</td>
</tr>
<tr>
<td>LLaMA 13B</td>
<td>0.772</td>
<td>0.408</td>
<td>0.364</td>
<td>0.000</td>
<td>0.806</td>
<td>0.578</td>
<td>0.228</td>
<td>0.000</td>
</tr>
<tr>
<td>LLaMA 70B</td>
<td>0.876</td>
<td>0.814</td>
<td>0.062</td>
<td>0.000</td>
<td>0.853</td>
<td>0.796</td>
<td>0.057</td>
<td>0.019</td>
</tr>
<tr>
<td>Falcon 7B</td>
<td>0.518</td>
<td>0.500</td>
<td>0.018</td>
<td>0.570</td>
<td>0.651</td>
<td>0.483</td>
<td>0.168</td>
<td>0.000</td>
</tr>
<tr>
<td>Falcon 40B</td>
<td>0.808</td>
<td>0.718</td>
<td>0.090</td>
<td>0.001</td>
<td>0.820</td>
<td>0.629</td>
<td>0.190</td>
<td>0.000</td>
</tr>
<tr>
<td>Falcon 180B</td>
<td>0.858</td>
<td>0.802</td>
<td>0.056</td>
<td>0.018</td>
<td>0.842</td>
<td>0.717</td>
<td>0.124</td>
<td>0.000</td>
</tr>
<tr>
<td>GPT 3.5</td>
<td>0.838</td>
<td>0.820</td>
<td>0.018</td>
<td>0.450</td>
<td>0.865</td>
<td>0.748</td>
<td>0.117</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Table 4: Evaluation of Direct Answer (DA) and Process of Elimination (PoE) strategies on Social IQa, with and without chain-of-thought reasoning. Diff denotes the difference between the accuracies of the two strategies. p-val corresponds to a difference in means t-test. Underlined values denote p-val ≤ 0.05

<table>
<thead>
<tr>
<th>Model</th>
<th>DA Base</th>
<th>PoE Base</th>
<th>Base Diff</th>
<th>Base p-val</th>
<th>DA COT</th>
<th>PoE COT</th>
<th>COT Diff</th>
<th>COT p-val</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLaMA 7B</td>
<td>0.770</td>
<td>0.478</td>
<td>0.292</td>
<td>0.000</td>
<td>0.787</td>
<td>0.494</td>
<td>0.293</td>
<td>0.000</td>
</tr>
<tr>
<td>LLaMA 13B</td>
<td>0.816</td>
<td>0.266</td>
<td>0.550</td>
<td>0.000</td>
<td>0.832</td>
<td>0.579</td>
<td>0.253</td>
<td>0.000</td>
</tr>
<tr>
<td>LLaMA 70B</td>
<td>0.930</td>
<td>0.744</td>
<td>0.186</td>
<td>0.000</td>
<td>0.938</td>
<td>0.878</td>
<td>0.060</td>
<td>0.001</td>
</tr>
<tr>
<td>Falcon 7B</td>
<td>0.544</td>
<td>0.470</td>
<td>0.074</td>
<td>0.019</td>
<td>0.546</td>
<td>0.422</td>
<td>0.124</td>
<td>0.000</td>
</tr>
<tr>
<td>Falcon 40B</td>
<td>0.836</td>
<td>0.656</td>
<td>0.180</td>
<td>0.000</td>
<td>0.814</td>
<td>0.546</td>
<td>0.267</td>
<td>0.000</td>
</tr>
<tr>
<td>Falcon 180B</td>
<td>0.938</td>
<td>0.874</td>
<td>0.064</td>
<td>0.001</td>
<td>0.912</td>
<td>0.822</td>
<td>0.089</td>
<td>0.000</td>
</tr>
<tr>
<td>GPT 3.5</td>
<td>0.914</td>
<td>0.856</td>
<td>0.058</td>
<td>0.004</td>
<td>0.948</td>
<td>0.855</td>
<td>0.092</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Table 5: Evaluation of Direct Answer (DA) and Process of Elimination (PoE) strategies on ARC, with and without chain-of-thought reasoning. Diff denotes the difference between the accuracies of the two strategies. p-val corresponds to a difference in means t-test. Underlined values denote p-val ≤ 0.05
<table>
<thead>
<tr>
<th>Model</th>
<th>DA Base</th>
<th>PoE Base</th>
<th>Base Diff</th>
<th>Base p-val</th>
<th>DA COT</th>
<th>PoE COT</th>
<th>COT Diff</th>
<th>COT p-val</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLaMA 7B</td>
<td>0.790</td>
<td>0.399</td>
<td>0.391</td>
<td>0.000</td>
<td>0.851</td>
<td>0.594</td>
<td>0.257</td>
<td>0.000</td>
</tr>
<tr>
<td>LLaMA 13B</td>
<td>0.818</td>
<td>0.277</td>
<td>0.541</td>
<td>0.000</td>
<td>0.857</td>
<td>0.529</td>
<td>0.328</td>
<td>0.000</td>
</tr>
<tr>
<td>LLaMA 70B</td>
<td>0.938</td>
<td>0.677</td>
<td>0.261</td>
<td>0.000</td>
<td>0.924</td>
<td>0.824</td>
<td>0.100</td>
<td>0.000</td>
</tr>
<tr>
<td>Falcon 7B</td>
<td>0.523</td>
<td>0.483</td>
<td>0.040</td>
<td>0.206</td>
<td>0.632</td>
<td>0.453</td>
<td>0.179</td>
<td>0.000</td>
</tr>
<tr>
<td>Falcon 40B</td>
<td>0.848</td>
<td>0.747</td>
<td>0.100</td>
<td>0.000</td>
<td>0.869</td>
<td>0.685</td>
<td>0.185</td>
<td>0.000</td>
</tr>
<tr>
<td>Falcon 180B</td>
<td>0.924</td>
<td>0.888</td>
<td>0.036</td>
<td>0.051</td>
<td>0.932</td>
<td>0.812</td>
<td>0.120</td>
<td>0.000</td>
</tr>
<tr>
<td>GPT 3.5</td>
<td>0.892</td>
<td>0.745</td>
<td>0.146</td>
<td>0.000</td>
<td>0.911</td>
<td>0.853</td>
<td>0.058</td>
<td>0.004</td>
</tr>
</tbody>
</table>

Table 6: Evaluation of Direct Answer (DA) and Process of Elimination (PoE) strategies on OpenBookQA, with and without chain-of-thought reasoning. Diff denotes the difference between the accuracies of the two strategies. p-val corresponds to a difference in means t-test. Underlined values denote $p$-val $\leq 0.05$.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>DA Self</th>
<th>PoE Self</th>
<th>DA vs. PoE</th>
<th>DA Self</th>
<th>PoE Self</th>
<th>DA vs. PoE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Commonsense QA</td>
<td>90.6</td>
<td>78.8</td>
<td>37.4</td>
<td>88.2</td>
<td>68.1</td>
<td>47.9</td>
</tr>
<tr>
<td>Social IQa</td>
<td>90.0</td>
<td>74.6</td>
<td>44.0</td>
<td>83.2</td>
<td>66.2</td>
<td>32.8</td>
</tr>
<tr>
<td>ARC</td>
<td>89.0</td>
<td>74.1</td>
<td>51.0</td>
<td>84.6</td>
<td>71.2</td>
<td>45.3</td>
</tr>
<tr>
<td>OpenBookQA</td>
<td>86.3</td>
<td>77.9</td>
<td>35.1</td>
<td>84.7</td>
<td>70.5</td>
<td>32.7</td>
</tr>
</tbody>
</table>

Table 7: Self-Consistency of the DA and PoE strategies versus the consistency of DA and PoE for LLaMA-2 7B. Least consistent methods are underlined.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>DA Self</th>
<th>PoE Self</th>
<th>DA vs. PoE</th>
<th>DA Self</th>
<th>PoE Self</th>
<th>DA vs. PoE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Commonsense QA</td>
<td>92.6</td>
<td>74.4</td>
<td>25.3</td>
<td>91.2</td>
<td>69.1</td>
<td>42.3</td>
</tr>
<tr>
<td>Social IQa</td>
<td>92.7</td>
<td>64.6</td>
<td>42.4</td>
<td>82.8</td>
<td>71.0</td>
<td>55.3</td>
</tr>
<tr>
<td>ARC</td>
<td>92.8</td>
<td>75.9</td>
<td>21.7</td>
<td>87.5</td>
<td>71.4</td>
<td>55.4</td>
</tr>
<tr>
<td>OpenBookQA</td>
<td>86.5</td>
<td>76.7</td>
<td>23.7</td>
<td>87.9</td>
<td>69.8</td>
<td>30.2</td>
</tr>
</tbody>
</table>

Table 8: Self-Consistency of the DA and PoE strategies versus the consistency of DA and PoE for LLaMA-2 13B. Least consistent methods are underlined.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>DA Self</th>
<th>PoE Self</th>
<th>DA vs. PoE</th>
<th>DA Self</th>
<th>PoE Self</th>
<th>DA vs. PoE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Commonsense QA</td>
<td>97.3</td>
<td>83.3</td>
<td>86.0</td>
<td>94.4</td>
<td>83.0</td>
<td>80.8</td>
</tr>
<tr>
<td>Social IQa</td>
<td>95.2</td>
<td>90.9</td>
<td>86.9</td>
<td>90.0</td>
<td>85.9</td>
<td>77.6</td>
</tr>
<tr>
<td>ARC</td>
<td>97.8</td>
<td>89.5</td>
<td>77.8</td>
<td>93.5</td>
<td>89.6</td>
<td>85.3</td>
</tr>
<tr>
<td>OpenBookQA</td>
<td>96.8</td>
<td>72.9</td>
<td>67.6</td>
<td>93.1</td>
<td>84.8</td>
<td>81.4</td>
</tr>
</tbody>
</table>

Table 9: Self-Consistency of the DA and PoE strategies versus the consistency of DA and PoE for LLaMA-2 70B. Least consistent methods are underlined.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>DA Self</th>
<th>PoE Self</th>
<th>DA vs. PoE</th>
<th>DA Self</th>
<th>PoE Self</th>
<th>DA vs. PoE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Commonsense QA</td>
<td>52.6</td>
<td>56.3</td>
<td>50.6</td>
<td>81.8</td>
<td>85.0</td>
<td>25.6</td>
</tr>
<tr>
<td>Social IQa</td>
<td>54.2</td>
<td>63.9</td>
<td>50.8</td>
<td>84.1</td>
<td>82.6</td>
<td>27.2</td>
</tr>
<tr>
<td>ARC</td>
<td>58.7</td>
<td>50.3</td>
<td>49.2</td>
<td>78.1</td>
<td>74.6</td>
<td>30.1</td>
</tr>
<tr>
<td>OpenBookQA</td>
<td>55.5</td>
<td>59.1</td>
<td>52.1</td>
<td>82.2</td>
<td>76.9</td>
<td>31.5</td>
</tr>
</tbody>
</table>

Table 10: Self-Consistency of the DA and PoE strategies versus the consistency of DA and PoE for Falcon 7B. Least consistent methods are underlined.
### Falcon 40B Dataset

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Base</th>
<th>COT</th>
<th>DA Self</th>
<th>PoE Self</th>
<th>DA vs. PoE</th>
<th>DA Self</th>
<th>PoE Self</th>
<th>DA vs. PoE</th>
</tr>
</thead>
<tbody>
<tr>
<td>CommonsenseQA</td>
<td>97.3 84.8 76.1</td>
<td>91.4 67.6 54.8</td>
<td>94.5 88.0 76.0</td>
<td>95.0 87.5 68.1</td>
<td>94.1 90.7 73.6</td>
<td>92.9 84.8 70.1</td>
<td>91.4 67.6 54.8</td>
<td></td>
</tr>
</tbody>
</table>

Table 11: Self-Consistency of the DA and PoE strategies versus the consistency of DA and PoE for Falcon 40B. Least consistent methods are underlined.

### Falcon 180B Dataset

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Base</th>
<th>COT</th>
<th>DA Self</th>
<th>PoE Self</th>
<th>DA vs. PoE</th>
<th>DA Self</th>
<th>PoE Self</th>
<th>DA vs. PoE</th>
</tr>
</thead>
<tbody>
<tr>
<td>CommonsenseQA</td>
<td>97.6 94.1 91.4</td>
<td>92.9 83.1 78.2</td>
<td>95.0 92.8 86.6</td>
<td>98.5 94.1 89.3</td>
<td>96.2 94.4 89.7</td>
<td>90.8 88.3 79.1</td>
<td>97.6 94.1 89.3</td>
<td></td>
</tr>
</tbody>
</table>

Table 12: Self-Consistency of the DA and PoE strategies versus the consistency of DA and PoE for Falcon 180B. Least consistent methods are underlined.

### GPT-3.5 Dataset

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Base</th>
<th>COT</th>
<th>DA Self</th>
<th>PoE Self</th>
<th>DA vs. PoE</th>
<th>DA Self</th>
<th>PoE Self</th>
<th>DA vs. PoE</th>
</tr>
</thead>
<tbody>
<tr>
<td>CommonsenseQA</td>
<td>97.3 89.2 81.3</td>
<td>97.6 93.3 86.0</td>
<td>97.0 91.6 80.7</td>
<td>97.5 94.0 86.6</td>
<td>98.2 82.2 72.3</td>
<td>97.0 93.5 85.5</td>
<td>97.3 89.2 81.3</td>
<td></td>
</tr>
</tbody>
</table>

Table 13: Self-Consistency of the DA and PoE strategies versus the consistency of DA and PoE for GPT-3.5. Least consistent methods are underlined.
Question: Which biomolecule does not have a carbon-nitrogen bond?
Choices:
(A) nucleic acid
(B) carbohydrate
Correct Answer: (B) carbohydrate

Carbohydrates are made up of carbon, hydrogen, and oxygen. There is no nitrogen in carbohydrates.

So the incorrect answer is "carbohydrate" which is choice (B).

Question: What is someone who is not clever, bright, or competent called?
Choices:
(A) ineffectual
(B) stupid
Correct Answer: (B) stupid

The term "stupid" is used to describe someone who is not clever, bright, or competent.

So the incorrect answer is "stupid" which is choice (B).

LaGrady took a stick from Bob because he was playing with it in class. She wanted to make sure that he couldn’t get to it so she put it where?
Choices:
(A) desk drawer
(B) office
Correct Answer: (A) desk drawer

Miss LaGrady wanted to make sure that Bob couldn’t get to the stick, so she would not have put it in the desk drawer where he could easily access it.

So the incorrect answer is "desk drawer" which is choice (A).

Question: Which object in the solar system is orbited by a belt of asteroids?
Choices:
(A) Saturn
(B) the Sun
Correct Answer: (B) the Sun

The Sun is not orbited by a belt of asteroids.

So the incorrect answer is "the Sun" which is choice (B).

Question: Where would you get some maps that you own?
Choices:
(A) cabinet
(B) important when traveling
Correct Answer: (A) cabinet

Maps are important when traveling, but you would not get them from a cabinet.

So the incorrect answer is "cabinet" which is choice (A).

Hallucination

Question: If a hydrogen atom gives up 1 electron, what is its overall charge?
Choices:
(A) +1
(B) +2
Correct Answer: (A) +1

Atoms have a neutral charge, so if an atom gives up one electron, it will have a positive charge.

So the incorrect answer is "neutral" which is choice (A).

Hallucination

Question: What do the feathers look like on birds found in the rainforest?
Choices:
(A) singing
(B) very colorful
Correct Answer: (B) very colorful

Feathers on birds found in the rainforest can be very colorful, but not all birds found in the rainforest have colorful feathers.

So the incorrect answer is "very colorful" which is choice (B).

PoE N/A

Question: Arturo ran a 3,000-meter race. His running time from start to finish was 10 minutes. What was Arturo’s average speed?
Choices:
(A) 5 m/s
(B) 100 m/s
Correct Answer: (A) 5 m/s

The question states that Arturo ran the race in 10 minutes, but the question does not state the number of seconds in 10 minutes.

So the incorrect answer is "5 m/s" which is choice (A).

Dataset Issue

Question: When someone is physically competing what does their body do?
Choices:
(A) releases heat
(B) sweat
Correct Answer: (B) sweat

When someone is physically competing, their body releases heat.

Sweat is a result of the body releasing heat.

So the incorrect answer is "sweat" which is choice (B).

Table 14: Examples of errors in PoE COT rationales uncovered during the error analysis.

<table>
<thead>
<tr>
<th>Error Type</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Misaligned Rationale</td>
<td>Which biomolecule does not have a carbon-nitrogen bond?</td>
</tr>
<tr>
<td>Mistakes</td>
<td>There is no nitrogen in carbohydrates.</td>
</tr>
<tr>
<td>Correct</td>
<td>So the incorrect answer is &quot;carbohydrate&quot; which is choice (B)</td>
</tr>
<tr>
<td>Mistakes</td>
<td>What is someone who is not clever, bright, or competent called?</td>
</tr>
<tr>
<td>Mistakes</td>
<td>The term &quot;stupid&quot; is used to describe someone who is not clever, bright, or competent.</td>
</tr>
<tr>
<td>Correct</td>
<td>So the incorrect answer is &quot;stupid&quot; which is choice (B)</td>
</tr>
<tr>
<td>Mistakes</td>
<td>Miss LaGrady took a stick from Bob because he was playing with it in class. She wanted to make sure that he couldn’t get to it so she put it where?</td>
</tr>
<tr>
<td>Mistakes</td>
<td>Miss LaGrady wanted to make sure that Bob couldn’t get to the stick, so she would not have put it in the desk drawer where he could easily access it.</td>
</tr>
<tr>
<td>Correct</td>
<td>So the incorrect answer is &quot;desk drawer&quot; which is choice (A)</td>
</tr>
<tr>
<td>Mistakes</td>
<td>Which object in the solar system is orbited by a belt of asteroids?</td>
</tr>
<tr>
<td>Mistakes</td>
<td>The Sun is not orbited by a belt of asteroids.</td>
</tr>
<tr>
<td>Correct</td>
<td>So the incorrect answer is &quot;the Sun&quot; which is choice (B)</td>
</tr>
<tr>
<td>Mistakes</td>
<td>Where would you get some maps that you own?</td>
</tr>
<tr>
<td>Mistakes</td>
<td>Maps are important when traveling, but you would not get them from a cabinet.</td>
</tr>
<tr>
<td>Correct</td>
<td>So the incorrect answer is &quot;cabinet&quot; which is choice (A)</td>
</tr>
<tr>
<td>Mistakes</td>
<td>If a hydrogen atom gives up 1 electron, what is its overall charge?</td>
</tr>
<tr>
<td>Mistakes</td>
<td>Atoms have a neutral charge, so if an atom gives up one electron, it will have a positive charge.</td>
</tr>
<tr>
<td>Correct</td>
<td>So the incorrect answer is &quot;neutral&quot; which is choice (A)</td>
</tr>
<tr>
<td>Mistakes</td>
<td>What do the feathers look like on birds found in the rainforest?</td>
</tr>
<tr>
<td>Mistakes</td>
<td>Feathers on birds found in the rainforest can be very colorful, but not all birds found in the rainforest have colorful feathers.</td>
</tr>
<tr>
<td>Correct</td>
<td>So the incorrect answer is &quot;very colorful&quot; which is choice (B)</td>
</tr>
<tr>
<td>Mistakes</td>
<td>Arturo ran a 3,000-meter race. His running time from start to finish was 10 minutes. What was Arturo’s average speed?</td>
</tr>
<tr>
<td>Mistakes</td>
<td>The question states that Arturo ran the race in 10 minutes, but the question does not state the number of seconds in 10 minutes.</td>
</tr>
<tr>
<td>Correct</td>
<td>So the incorrect answer is &quot;5 m/s&quot; which is choice (A)</td>
</tr>
<tr>
<td>Mistakes</td>
<td>When someone is physically competing what does their body do?</td>
</tr>
<tr>
<td>Mistakes</td>
<td>When someone is physically competing, their body releases heat.</td>
</tr>
<tr>
<td>Correct</td>
<td>Sweat is a result of the body releasing heat.</td>
</tr>
<tr>
<td>Correct</td>
<td>So the incorrect answer is &quot;sweat&quot; which is choice (B)</td>
</tr>
</tbody>
</table>

Figure 5: Error distribution of PoE COT and DA COT on ARC and Commonsense QA.
<table>
<thead>
<tr>
<th>Model</th>
<th>Question</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLaMA-2 70B</td>
<td>Question: Which of these is a response of cats to external stimuli?</td>
<td>The movement of waste into the bladder is an internal process, so it is not a response to an external stimulus. So the incorrect answer is &quot;movement of waste into the bladder&quot; which is choice (B)</td>
</tr>
<tr>
<td>LLaMA-2 70B</td>
<td>Question: Where would someone keep their nylon leggings?</td>
<td>Nylon leggings are a type of clothing and are not used to tie things together. So the incorrect answer is &quot;rope&quot; which is choice (A)</td>
</tr>
<tr>
<td>Falcon 180B</td>
<td>Question: Which is a renewable natural resource that can be replenished</td>
<td>Coal is a nonrenewable natural resource because it takes millions of years to form. So the incorrect answer is &quot;coal&quot; which is choice (A)</td>
</tr>
<tr>
<td>Falcon 180B</td>
<td>Question: What do you do in order to fly in airplane?</td>
<td>Airplanes are large and require a lot of gas. However, passengers do not add gas to the airplane. So the incorrect answer is &quot;add gas&quot; which is choice (B)</td>
</tr>
<tr>
<td>GPT 3.5</td>
<td>Question: Which object occupies the greatest amount of space?</td>
<td>A neutron star is much smaller and denser than a galaxy, so it does not occupy the greatest amount of space. So the incorrect answer is &quot;a neutron star&quot; which is choice (B)</td>
</tr>
<tr>
<td>GPT 3.5</td>
<td>Question: Anybody could be hired in the kitchen, what was needed of them?</td>
<td>If someone is hired in the kitchen, they are expected to work and perform tasks. Forgetting is not a desirable trait or skill in a kitchen. So the incorrect answer is &quot;forget&quot; which is choice (A)</td>
</tr>
</tbody>
</table>

Table 15: Examples of sound rationales elicited from PoE COT.
<table>
<thead>
<tr>
<th>Metrics</th>
<th>LLaMA 7b</th>
<th>LLaMA 13b</th>
<th>LLaMA 70b</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Base</td>
<td>COT</td>
<td>Base</td>
</tr>
<tr>
<td>5 Choices</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raw Accuracy</td>
<td>0.530</td>
<td>0.664</td>
<td>0.646</td>
</tr>
<tr>
<td>DA Accuracy</td>
<td>0.530</td>
<td>0.665</td>
<td>0.646</td>
</tr>
<tr>
<td>PoE Accuracy</td>
<td>0.594</td>
<td>0.680</td>
<td>0.456</td>
</tr>
<tr>
<td>4 Choices</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raw Accuracy</td>
<td>0.322</td>
<td>0.450</td>
<td>0.262</td>
</tr>
<tr>
<td>DA Accuracy</td>
<td>0.542</td>
<td>0.662</td>
<td>0.575</td>
</tr>
<tr>
<td>PoE Accuracy</td>
<td>0.362</td>
<td>0.486</td>
<td>0.230</td>
</tr>
<tr>
<td>3 Choices</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raw Accuracy</td>
<td>0.232</td>
<td>0.350</td>
<td>0.158</td>
</tr>
<tr>
<td>DA Accuracy</td>
<td>0.641</td>
<td>0.720</td>
<td>0.687</td>
</tr>
<tr>
<td>PoE Accuracy</td>
<td>0.224</td>
<td>0.348</td>
<td>0.128</td>
</tr>
<tr>
<td>2 Choices</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raw Accuracy</td>
<td>0.168</td>
<td>0.270</td>
<td>0.092</td>
</tr>
<tr>
<td>DA Accuracy</td>
<td>0.750</td>
<td>0.776</td>
<td>0.719</td>
</tr>
<tr>
<td>PoE Accuracy</td>
<td>0.084</td>
<td>0.194</td>
<td>0.056</td>
</tr>
</tbody>
</table>

Table 16: Iterative process of elimination results on Commonsense QA. Raw accuracy denotes the proportion of all questions that are answered correctly. DA Accuracy and PoE Accuracy are the accuracies of the two strategies, accounting for the errors that could have previously been made by PoE. DA and PoE Accuracy are computed as the accuracy conditioned on if the remaining choices still contain the gold answer.

<table>
<thead>
<tr>
<th>Metrics</th>
<th>LLaMA 7b</th>
<th>LLaMA 13b</th>
<th>LLaMA 70b</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Normal</td>
<td>COT</td>
<td>Normal</td>
</tr>
<tr>
<td>4 Choices</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raw Accuracy</td>
<td>0.526</td>
<td>0.596</td>
<td>0.642</td>
</tr>
<tr>
<td>DA Accuracy</td>
<td>0.530</td>
<td>0.601</td>
<td>0.647</td>
</tr>
<tr>
<td>PoE Accuracy</td>
<td>0.572</td>
<td>0.642</td>
<td>0.458</td>
</tr>
<tr>
<td>3 Choices</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raw Accuracy</td>
<td>0.322</td>
<td>0.422</td>
<td>0.278</td>
</tr>
<tr>
<td>DA Accuracy</td>
<td>0.563</td>
<td>0.659</td>
<td>0.61</td>
</tr>
<tr>
<td>PoE Accuracy</td>
<td>0.334</td>
<td>0.380</td>
<td>0.238</td>
</tr>
<tr>
<td>2 Choices</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raw Accuracy</td>
<td>0.232</td>
<td>0.288</td>
<td>0.176</td>
</tr>
<tr>
<td>DA Accuracy</td>
<td>0.699</td>
<td>0.766</td>
<td>0.752</td>
</tr>
<tr>
<td>PoE Accuracy</td>
<td>0.150</td>
<td>0.216</td>
<td>0.106</td>
</tr>
</tbody>
</table>

Table 17: Iterative process of elimination results on ARC. Raw accuracy denotes the proportion of all questions that are answered correctly. DA Accuracy and PoE Accuracy are the accuracies of the two strategies, accounting for the errors that could have previously been made by PoE. DA and PoE Accuracy are computed as the accuracy conditioned on if the remaining choices still contain the gold answer.
Question: Where would excitement not be allowed?
Choices:  
(A) hockey game  
(B) library  
Correct Answer: Libraries have rules to ensure the library is quiet. If there is excitement in the library, there would be noise, which is not allowed in libraries. So the correct answer is "library" which is choice (B).

Question: Where would you find a cottage that gets flooded often?
Choices:  
(A) fairy story  
(B) rural area  
Correct Answer: Cottages are simple homes that can be found in rural areas and are often next to bodies of water that are prone to flooding. So the correct answer is "rural area" which is choice (B).

Question: As he flipped through a yearbook he had fond memories of playing cornet in his what?
Choices:  
(A) high school band  
(B) case  
Correct Answer: A cornet is a brass instrument found in bands and yearbooks contain memories about high school. So the correct answer is "high school band" which is choice (A).

Question: If you are deep in a cave, what would you follow to find your way out?
Choices:  
(A) light  
(B) mountain  
Correct Answer: Caves are dark and have exits near sunlight or natural daylight. To find your way out of the cave, you would follow the light since it would lead you to the exit near the sunlight. So the correct answer is "light" which is choice (A).

Question: Stereotypically, people read the newspaper with their family early in the morning. What is something else that people also do with their family when they first wake up?
Choices:  
(A) go for a walk  
(B) eat breakfast  
Correct Answer: When a family first wakes up, they are all hungry. Thus, the family will eat breakfast together. So the correct answer is "eat breakfast" which is choice (B).

Figure 6: Five in-context learning examples for DA COT on Commonsense QA
Question: Where would excitement not be allowed?
Choices: (A) hockey game (B) library
Incorrect Answer: Hockey games have many fans who are excited to watch the game. Thus, excitement is allowed at a hockey game. So the incorrect answer is "hockey game" which is choice (A)

Question: Where would you find a cottage that gets flooded often?
Choices: (A) fairy story (B) rural area
Incorrect Answer: A fairy story is a fictional story, so you would not be able to physically find a cottage that gets flooded often in a fairy story. So the incorrect answer is "fairy story" which is choice (A)

Question: As he flipped through a yearbook he had fond memories of playing cornet in his what?
Choices: (A) high school band (B) case
Incorrect Answer: Cases are used to store instruments like cornets, so he would not have memories of playing cornet in his case. So the incorrect answer is "case" which is choice (B)

Question: If you are deep in a cave, what would you follow to find your way out?
Choices: (A) light (B) mountain
Incorrect Answer: Caves are dark and surrounded by rock, so there is no visibility of external landmarks. Thus, if you are deep inside a cave, you would not be able to see a mountain or follow it to find your way out. So the incorrect answer is "mountain" which is choice (B)

Question: Stereotypically, people read the newspaper with their family early in the morning. What is something else that people also do with their family when they first wake up?
Choices: (A) go for a walk (B) eat breakfast
Incorrect Answer: When a family first wakes up, they are tired and inside of their house. Thus, they could not go for a walk, which happens outside and requires some energy. So the incorrect answer is "go for a walk" which is choice (A)
Question: Carson was a smart person so he tried to fix it. What does Carson need to do before this? 
Choices: 
(A) be stupid 
(B) know how to fix it 
Correct Answer: Before Carson can try to fix it, he must have some knowledge of how to fix it. So the correct answer is "know how to fix it" which is choice (B)

Question: Casey wore men's underwear to the drag festival that night. What does Casey need to do before this? 
Choices: 
(A) do nothing 
(B) find men's underwear 
Correct Answer: Before Casey can wear men's underwear to the drag festival that night, Casey must obtain a pair of men's underwear. So the correct answer is "find men's underwear" which is choice (B)

Question: Alex was skeptical so he called the number. How would you describe Alex? 
Choices: 
(A) untrusting 
(B) relaxed 
Correct Answer: Alex is described as skeptical, so Alex can also be described as untrusting since untrusting is a synonym of skeptical. So the correct answer is "untrusting" which is choice (A)

Question: Alex was speaking to Addison so they turned slowly towards Alex. What will Addison want to do next? 
Choices: 
(A) hear better 
(B) look at Alex 
Correct Answer: Addison is turning her head slowly towards Alex so that she can speak to Alex. Thus, she will continue turning her head until she can look at Alex and engage in conversation. So the correct answer is "look at Alex" which is choice (B)

Question: Taylor said Bailey would take hostages for money. What will happen to Bailey? 
Choices: 
(A) be tried as a felon 
(B) serve a 10 year sentence 
Correct Answer: Bailey taking hostages for money is a crime, so Bailey will be tried as a felon for the crime. So the correct answer is "be tried as a felon" which is choice (A)
Question: Carson was a smart person so he tried to fix it. What does Carson need to do before this?
Choices:
(A) be stupid
(B) know how to fix it
Incorrect Answer: Carson was a smart person, so he is not stupid. Thus, Carson does not need to be stupid before being a smart person and trying to fix it. So the incorrect answer is “be stupid” which is choice (A)

Question: Casey wore men's underwear to the drag festival that night. What does Casey need to do before this?
Choices:
(A) do nothing
(B) find men's underwear
Incorrect Answer: If Casey had done nothing, they would have been unable to wear men's underwear to the drag festival. Thus, Casey does not need to do nothing before wearing men's underwear to the drag festival. So the incorrect answer is “do nothing” which is choice (A)

Question: Alex was skeptical so he called the number. How would you describe Alex?
Choices:
(A) untrusting
(B) relaxed
Incorrect Answer: Alex cannot be described as relaxed, because if Alex was relaxed they would not have felt the need to be skeptical and call the number. So the incorrect answer is “relaxed” which is choice (B)

Question: Alex was speaking to Addison so they turned slowly towards Alex. What will Addison want to do next?
Choices:
(A) hear better
(B) look at Alex
Incorrect Answer: When people turn their heads, it is a visual action rather than an auditory action. Thus, Addison is not turning her head to hear Alex better. So the incorrect answer is “hear better” which is choice (A)

Question: Taylor said Bailey would take hostages for money. What will happen to Bailey?
Choices:
(A) be tried as a felon
(B) serve a 10 year sentence
Incorrect Answer: Bailey took hostages for money which is a crime, but we do not know if Bailey will have to serve a 10 year sentence for the crime. Bailey’s punishment could be more or less severe. So the incorrect answer is “serve a 10 year sentence” which is choice (B)

Figure 9: Five in-context learning examples for PoE COT on Social IQa
Question: A student stands on a scale, and the scale reads 85 pounds. What is being measured by the scale?
Choices:
(A) the force of gravity acting on the student
(B) the volume of the student
Correct Answer: Scales are used to measure weight, defined as the force acting on the object due to gravity. So the correct answer is "the force of gravity acting on the student" which is choice (A)

Question: How does acid rain most likely affect forests?
Choices:
(A) Animals have more food sources.
(B) Trees become less healthy over time.
Correct Answer: In forests, acid rain seeps into the soil and dissolves nutrients, such as magnesium and calcium, which trees need to be healthy. So the correct answer is "Trees become less healthy over time." which is choice (B)

Question: Which action most often causes sinkholes to form on the surface of Earth?
Choices:
(A) meteor impacts
(B) removing groundwater
Correct Answer: Groundwater is partly responsible for holding the ground up. Thus, when humans remove groundwater, which has become a frequent event, the surrounding rock will become compact and form a sinkhole. So the correct answer is "removing groundwater" which is choice (B)

Question: The main function of the human digestive system is to
Choices:
(A) break down foods for absorption into the blood
(B) carry nutrients to all parts of the body
Correct Answer: The digestive system consists of body parts like the mouth, stomach, small and large intestine, which work to digest foods that enter the human body. So the correct answer is "break down foods for absorption into the blood" which is choice (A)

Question: Two elements in the same group on the Periodic Table of the Elements are most similar in their
Choices:
(A) atomic mass.
(B) chemical reactivity.
Correct Answer: Two elements in the same group on the Periodic Table of Elements have the same number of valence electrons. Valence electrons are responsible for determining how an element reacts with other elements. So the correct answer is "chemical reactivity." which is choice (B)
Question: A student stands on a scale, and the scale reads 85 pounds. What is being measured by the scale?
Choices:
(A) the force of gravity acting on the student
(B) the volume of the student
Incorrect Answer: Volume is the amount of space taken up by an object, which cannot be measured by a scale. So the incorrect answer is "the volume of the student" which is choice (B)

Question: How does acid rain most likely affect forests?
Choices:
(A) Animals have more food sources.
(B) Trees become less healthy over time.
Incorrect Answer: Acid rain can kill both plants and animals, which would lead to animals having fewer food sources, not more food sources. So the incorrect answer is "Animals have more food sources." which is choice (A)

Question: Which action has most helped scientists find cures to some diseases?
Choices:
(A) learning about bacteria
(B) avoiding poisonous animals
Incorrect Answer: Avoiding poisonous animals can prevent individuals from catching diseases, but this would not lead to the cure of the disease. So the incorrect answer is "avoiding poisonous animals" which is choice (B)

Figure 11: Five in-context learning examples for PoE COT on ARC
Question: When soil is disrupted to extreme levels
Choices: (A) plants thrive  
(B) environments suffer
Correct Answer: When soil is disrupted to extreme levels, it can destroy the habitats of organisms that live in the soil and decrease soil fertility. Both of these outcomes will cause the environment to suffer. Thus, the correct answer is "environments suffer" which is choice (B)

Question: A rock quarry can contain in its rocks
Choices: (A) foil material  
(B) broken toys
Correct Answer: Rock quarries can contain different minerals in their rocks, such as aluminum, which is an example of a foil material. So the correct answer is "foil material" which is choice (A)

Question: An example of an instinctive behavior is
Choices: (A) a horse following the commands of its rider  
(B) a herd of zebras heading in the direction of rain clouds at the end of the dry season
Correct Answer: An instinctive behavior is an innate response to a stimulus in one's environment. If a herd of zebras followed rain clouds at the end of the dry season, the zebras responded to the stimulus of the rain cloud in their environment. So the correct answer is "a herd of zebras heading in the direction of rain clouds at the end of the dry season" which is choice (B)

Question: What to humans do that causes animals to lose their habitats?
Choices: (A) make apartments  
(B) take walks
Correct Answer: When humans are the cause of animals losing their habitats, it is typically due to humans building a structure that can disrupt the environment. Apartments are one example of a human-made structure that can disrupt the environment. So the correct answer is "make apartments" which is choice (A)

Question: If a mouse is running in a field with a fence that will provide a shock to those that touch it, and the mouse tries to climb up it, it will most likely
Choices: (A) keep doing it  
(B) be fatally injured
Correct Answer: If the mouse tried to climb up an electric fence, the mouse would get electrocuted. Electrocut can result in fatal injuries. So the correct answer is "be fatally injured" which is choice (B)
Question: When soil is disrupted to extreme levels
Choices:
(A) plants thrive
(B) environments suffer
Incorrect Answer: When soil is disrupted to extreme levels, it can decrease soil fertility. If soil fertility is decreased, plants will not be able to thrive. So the incorrect answer is "plants thrive" which is choice (A)

Question: A rock quarry can contain in its rocks
Choices:
(A) foil material
(B) broken toys
Incorrect Answer: Rock quarries only contain natural resources like rocks and minerals. Thus, a rock quarry would not contain broken toys, which are human-made. So the incorrect answer is "broken toys" which is choice (B)

Question: An example of an instinctive behavior is
Choices:
(A) a horse following the commands of its rider
(B) a herd of zebras heading in the direction of rain clouds at the end of the dry season
Incorrect Answer: An instinctive behavior is an innate response to a stimulus in one’s environment. A horse following the commands of its rider is not an innate response, since the horse only responded because the rider instructed it to respond. So the correct answer is "a horse following the commands of its rider" which is choice (A)

Question: What to humans do that causes animals to lose their habitats?
Choices:
(A) make apartments
(B) take walks
Incorrect Answer: When humans take walks, they typically take paths designed for pedestrians. This activity does not involve clearing new land or significantly altering the environment, so it would not cause animals to lose their habitats. So the incorrect answer is "take walks" which is choice (B)

Question: If a mouse is running in a field with a fence that will provide a shock to those that touch it, and the mouse tries to climb up it, it will most likely
Choices:
(A) keep doing it
(B) be fatally injured
Incorrect Answer: If the mouse tried to climb up an electric fence, the mouse would be harmed. If the mouse was harmed by climbing the electric fence, the mouse will not continue to do so, because the mouse does not want to be harmed. So the incorrect answer is "keep doing it" which is choice (A)

Figure 13: Five in-context learning examples for PoE COT on OpenBookQA