

Large Language Models Can Self-Correct with Minimal Effort

Anonymous ACL submission

Abstract

Intrinsic self-correct was a method that instructed large language models (LLMs) to verify and correct their responses without external feedback. Unfortunately, the study concluded that the LLMs could not self-correct reasoning yet. We find that a simple yet effective verification method can unleash inherent capabilities of the LLMs. That is to mask a key condition in the question, add the current response to construct a verification question, and predict the condition to verify the response. The condition can be an entity in an open-domain question or a numeric value in a math question, which requires minimal effort (via prompting) to identify. We propose an iterative verify-then-correct framework to progressively identify and correct (probably) false responses, named PROCo. We conduct experiments on three reasoning tasks. On average, PROCo, with GPT-3.5-Turbo-1106 as the backend LLM, yields +6.8 exact match on four open-domain question answering datasets, +14.1 accuracy on three arithmetic reasoning datasets, and +9.6 accuracy on a commonsense reasoning dataset, compared to Self-Correct.

1 Introduction

Reasoning is a cognitive process that uses evidence, arguments, and logic to arrive at conclusions or judgements (Huang and Chang, 2023). People have been exploiting and improving the reasoning ability of large language models (LLMs). Wei et al. proposed chain-of-thought (CoT) prompting and yielded promising results on several reasoning tasks, such as arithmetic reasoning (Kojima et al., 2022; Zhou et al., 2023), commonsense reasoning (Wei et al., 2022; Zhang et al., 2023; Wang et al., 2023b), and open-domain question answering (Wang et al., 2023a), using only a few or no reasoning exemplars. CoT guides LLMs to generate intermediate reasoning paths instead of generating

Method	NQ	CSQA	AQuA
CoT	40.3	72.9	51.3
Self-Correct	40.1	65.9	48.7
PROCo (Ours)	48.0	75.5	65.2

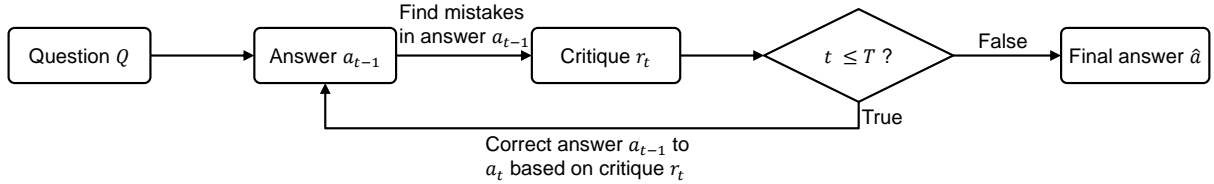
Table 1: Performance comparison of different prompting methods using GPT-3.5-Turbo as backend LLM.

the final answer directly, which helps the LLMs simulate the human-like reasoning process.

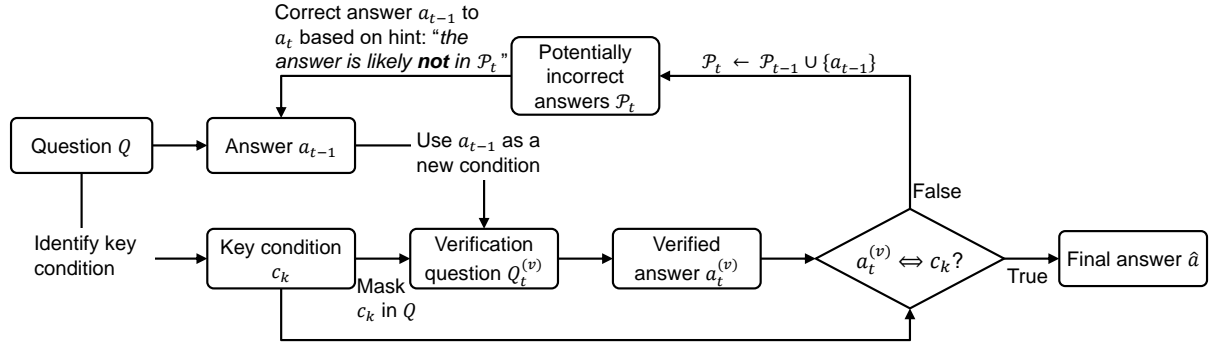
Although CoT enables LLMs to handle complex reasoning tasks, they are sensitive to mistakes in the reasoning path, as any mistake can lead to an incorrect answer. To address this issue, Dhuliawala et al.; Kim et al. have explored the verification and correction of responses. For example, as shown in Figure 1a, for a given question and its initial LLM-generated answer, Self-Correct (Kim et al., 2023) first instructs the LLM to criticize its generated answer using the hint: “Review previous answer and find mistakes”. Then, Self-Correct instructs the LLM to refine initial answers based on the critique.

However, recent studies (Huang et al., 2024; Gou et al., 2024) have cast doubt on the intrinsic self-correction capability of LLMs. Their research indicates that *without external feedback*, such as input from humans, other models, or external tools to verify the correctness of previous responses, LLMs struggle to correct their prior outputs. Since LLMs could not properly judge the correctness of their prior responses, the refined response might be even worse than the initial response.

To unleash inherent capabilities of LLMs to detect and rectify incorrect responses without external feedback, we introduce *substitute verification* (Yu et al., 2024). Let us look at a specific example. Given an open-domain question “Who plays Skylar on Lab Rats: Elite Force?”, we first prompt an LLM to generate an initial answer for the question, e.g., “Paris Berelc”. Next, we identify a key condi-



(a) Kim et al. proposed Self-Correct, instructing the LLM to critique and revise its answers using the hint “Review previous answer and find mistakes.” However, Huang et al. noted that LLMs struggle to correct mistakes without external feedback.



(b) PROCO performs three steps: (1) **Initialization**: Use CoT method to generate an initial answer. (2) **Verification**: Mask the key condition in the question and use the previous generated answer as a new condition to construct the verification question. Solve the verification question to get the verified answer and check if the verified answer and the key condition are equivalent. If they are equivalent, the previous generated answer is adopted as the final answer, otherwise add it to the set of potentially incorrect answers. (3) **Correction**: Use the set of potentially incorrect answers as feedback to correct previous generated answer. By cycle executing step (2) and step (3), the performance of LLMs on various complex reasoning tasks is progressively enhanced.

Figure 1: The proposed PROCO method helps LLMs identify incorrect answers and progressively correct them.

tion in the question that is relevant to the problem-solving process, such as “Skylar”. By masking the key condition in the question and adding the initial answer as a new condition, we can obtain a verification question: “Who plays X on Lab Rats: Elite Force? Suppose the answer is Paris Berelc. What is the value of unknown variable X?”. We use the LLM to solve the verification question, and we get that X is “Skylar Storm”. By verifying whether “Skylar Storm” is equivalent to “Skylar”, we can predict that the initial answer is likely correct.

Based on substitute verification, we propose a simple yet effective prompting method Progressive Correction (PROCO). Figure 1 illustrates the difference between the Self-Correct and PROCO methods. Compared with Self-Correct, our proposed PROCO highlights two primary distinctions:

(1) **Verification Method**. To improve verification accuracy, we propose the substitute verification method. Specifically, PROCO first identifies key conditions that are relevant to the problem-solving process. It then masks one of the key conditions in the question and takes the generated answer as a new condition to construct the verification question. Finally, PROCO solves the verification question and gets the verified answer. If the verified answer and

the key condition are equivalent, it indicates that the generated answer is likely to be correct.

(2) **Correction Method**. PROCO employs the substitute verification method to verify the correctness of LLM-generated answers. If an answer is deemed incorrect, PROCO adds it to a set of potentially incorrect answers, which then serves as feedback to guide LLMs in correcting previous mistakes with the hint: “the answer is likely not in {set of potentially incorrect answers}”. By iteratively executing verification and correction, PROCO prevents the repetition of previous mistakes, thereby progressively improving the quality of responses.

We conducted evaluations of PROCO using a variety of LLMs, including GPT-3.5-Turbo-1106, GPT-4-0125-Preview, and the open-source Mixtral-8x7B. These evaluations spanned three distinct tasks: arithmetic reasoning, commonsense reasoning, and open-domain question answering. The experimental results reveal that PROCO consistently outperforms existing methods. As shown in Table 1, PROCO achieves a 7.9 exact match (EM) improvement on the NQ dataset, a 16.5 absolute increase on the AQUA dataset, and a 9.6 absolute improvement on the CSQA dataset compared to the Self-Correct method.

In summary, our main contributions include:

- Based on our research, we have determined that LLMs are capable of intrinsic self-correction, provided that the prompt design is carefully structured within a framework focused on verification and correctness.
- We introduce a novel prompting method, PROCO, which utilizes an iterative verify-then-correct framework. PROCO progressively refines responses by identifying key conditions and formulating verification questions specific to these conditions.
- We conduct extensive experiments on three complex reasoning tasks and demonstrate that PROCO achieves significant improvements in both black-box and open-source LLMs.

2 Related Work

Self-Correct (Kim et al., 2023) methods, which aim to enhance the quality of LLM responses by providing feedback on initial attempts (Kim et al., 2023; Madaan et al., 2023; Chen et al., 2024), have demonstrated effectiveness in various reasoning tasks. These tasks include arithmetic reasoning (Madaan et al., 2023; Welleck et al., 2023), open-domain question answering (Dhuliawala et al., 2023; Yu et al., 2023b), commonsense reasoning (Kim et al., 2023), and others (Chen et al., 2024; Le et al., 2022). Self-Correct methods vary in the source and format of feedback, and the process of verifying the correctness of LLM output.

Source and Format of Feedback Inter-script (Tandon et al., 2021) corrected the LLM’s initial output by integrating natural language feedback from humans. Due to the high cost of human feedback, scalar reward functions have been used as alternatives. For instance, Rainer (Liu et al., 2022) used reinforcement learning to generate contextual relevant knowledge in response to queries. Self-Correction (Welleck et al., 2023) trained a corrector to iteratively correct imperfect outputs. Other sources, such as compilers (Chen et al., 2024) or search engines (Yu et al., 2023b) can provide external feedback.

Recent research used LLMs to generate feedback. Self-Correct (Kim et al., 2023) and Self-Refine (Madaan et al., 2023) utilized LLMs to verify and refine their initial outputs. However, Huang et al. questioned the intrinsic self-correcting capability of LLMs, indicating that without external feedback, LLMs struggle to correct their previ-

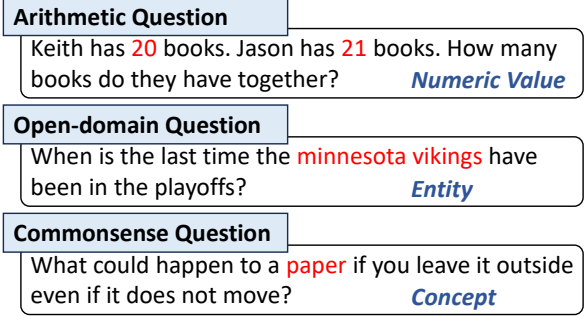


Figure 2: Key conditions in complex reasoning tasks play a crucial role in the problem-solving process. These conditions can take various forms: a numeric value in arithmetic questions, an entity in open-domain questions, or a concept in commonsense questions.

ous responses. To unleash the inherent capabilities of LLMs to detect and rectify incorrect responses without external feedback, we introduce *substitute verification*. By providing natural language feedback based on verification results, we can steer LLMs away from incorrect answers, thus enhancing their performance in various reasoning tasks.

Verify Correctness of LLM Output Several studies trained or fine-tuned language models to check the correctness of answers. Cobbe et al. fine-tuned GPT-3 as a verifier to judge the correctness of solutions. Li et al. fine-tuned DeBERTa-v3-large (He et al., 2021) to predict the probability that the generated reasoning path leads to a correct answer. Lightman et al. constructed a large dataset with step-wise correctness labels from human annotators, and fine-tuned a GPT-4 model on it. These methods require significant human annotations. To reduce human labor, Peng et al. proposed using an external database to identify incorrect knowledge in LLM outputs. Chern et al. used tools for fact-checking. Miao et al. used the LLM to verify the correctness of each step in the arithmetic reasoning path based on preceding steps. Dhuliawala et al. used manually crafted demonstrations as context to prompt the LLM to check the correctness of its output. All of these methods solely verify the correctness of LLM outputs and select the verified answer as the final answer. In contrast, our method iterates a verify-then-correct process to progressively identify and rectify incorrect answers.

3 Preliminaries

Given a question Q , consisting of m context sentences $\{s_j\}_{j=1}^m$ and one query sentence q . The

query q ends with a question mark and is usually the last sentence of Q . We can express $Q = (\oplus_j s_j) \oplus q$, where \oplus denotes text concatenation function. We extract conditions $\{c_i\}_{i=1}^n$ that are numerical values (arithmetic reasoning), entities (open-domain question answering), and concepts (commonsense reasoning), as shown in Figure 2. It is worth noting that usually $n \geq m$, if the question has one or multiple conditions. We denote $J(i) \in \{1, \dots, m\}$ as the index of the context sentence containing the condition c_i . Among these conditions, the key condition c_k is crucial for problem-solving and is used in the substitute verification process, where k is the index of the key condition within $\{c_i\}_{i=1}^n$. We introduce two innovative approaches for identifying the key condition.

Similarity-based Key Condition Identification

Numerical values are crucial in arithmetic reasoning tasks, so we select those relevant to solving the problem as key conditions. Key conditions are found in context sentences $\{s_j\}_{j=1}^m$ with high semantic relevance to the query sentence q . We use SimCSE (Gao et al., 2021) to encode the context and the query sentences, represented as $\{s_j\}_{j=1}^m$ and \mathbf{q} , respectively. Semantic relevance is calculated using cosine similarity between $\{s_j\}_{j=1}^m$ and \mathbf{q} . The most relevant context sentence index ℓ is determined by:

$$\ell = \operatorname{argmax}_{j \in \{1, \dots, m\}} \cos(s_j, \mathbf{q}). \quad (1)$$

We use regular expressions to extract the numerical value in context sentence s_ℓ as the key condition c_k . If multiple numerical values are present, one is randomly selected as the key condition.

Zero-shot Key Condition Identification Identifying key conditions in open-domain question answering (Entity) and commonsense reasoning (Concept) is not possible through regular expressions, unlike in arithmetic reasoning (Numerical Value). Instead, we directly instruct LLMs to identify these relevant entities or concepts as key conditions. For instance, given an open-domain question Q , we construct a key condition identification prompt:

“Given the question below, the task is to identify a set of entities within the question and then select the one that is most relevant to the problem-solving process. Q ”.

We then input this prompt into an LLM to obtain the key condition c_k .

4 Proposed Approach

4.1 Overview

In this section, we present the overall pipeline of the proposed Progressive Correction (PROCO) prompting method which consists of three steps. Figure 1b illustrates the PROCO method. Initially, PROCO prompts the LLM to generate an answer in response to a given question (Sec. 4.2). Subsequently, to enhance the preliminary answer, PROCO identifies a key condition and generates a corresponding verification question-answer pair based on that condition (Sec. 4.3). The final answer is refined by verifying the question-answer pair, ensuring the answer’s consistency and accuracy (Sec. 4.4). The full prompts used in the experiments can be found in Appendix A.4.

4.2 Generate Initial Answer

Given a question Q , we use one of the existing prompting methods, such as CoT (Kojima et al., 2022), RAG (Khatab et al., 2023), or GenRead (Yu et al., 2023a), to generate an initial answer a_0 . By default, we use the CoT (Kojima et al., 2022) prompting method to generate an initial answer.

4.3 Iterative Verify-then-Correct Process

We propose a novel iterative verify-then-correct method that first initializes the set of potentially incorrect answers as an empty set $\mathcal{P}_0 = \emptyset$ and identifies the key condition c_k within the question Q (Sec. 3). The method then progressively corrects the LLM-generated answer over T iterations by cyclically conducting verification and correction phases. Here we use the t -th iteration as an example to illustrate the verify-then-correct process.

Verification Phase The verification phase uses substitute verification method to verify the correctness of the previous generated answer a_{t-1} . This phase encompasses several substeps.

Initially, the key condition c_k within the question Q is replaced with a specific token “X”, resulting in a mask question:

$$Q^{(\text{mask})} = \left(\oplus_j s_j \Big|_{s_{J(k)} = s_{J(k)}^{(\text{mask})}} \right) \oplus q. \quad (2)$$

where $s_{J(k)}$ is the context sentence containing the key condition c_k , $s_{J(k)}^{(\text{mask})}$ denotes replacing c_k in $s_{J(k)}$ with “X”. We then construct the t -th verification question $Q_t^{(v)}$ based on the mask question:

$$Q_t^{(v)} = Q^{(\text{mask})} \oplus a_{t-1} \oplus q^{(v)} \quad (3)$$

where $q^{(v)}$ is a static question for verification, e.g., “What is the value of the unknown variable X ?” Note that through all iterations, the key condition remains the same, and we do not use it to construct $Q_t^{(v)}$, for any $t \in \{1, \dots, T\}$. The LLM is then instructed to solve the verification question $Q_t^{(v)}$ and produce the corresponding answer $a_t^{(v)}$. Finally, different strategies are proposed to verify the correctness of a_{t-1} .

Match-based Verification. For arithmetic questions, if $a_t^{(v)}$ is equal to c_k , it indicates that the previous answer a_{t-1} is most likely correct.

Proposition-based Verification. For open-domain or commonsense questions, we propose a proposition-based verification method to verify the correctness of the previously generated answer a_{t-1} . The intuition behind this is that the question $Q_t^{(v)}$ may have multiple valid answers, and directly checking if $a_t^{(v)}$ exactly matches c_k could result in misclassifying a correct answer as incorrect. Specifically, we construct an answer verification prompt: “Determine the correctness of the proposition: If the answer to question $Q_t^{(v)}$ is c_k , then X could also be $a_t^{(v)}$ ”. We input this prompt into an LLM and receive a judgment about the proposition’s correctness. If the proposition is verified as correct, it indicates that the previously generated answer a_{t-1} is likely correct, and we select a_{t-1} as the final answer \hat{a} and exit the loop. Otherwise, we add a_{t-1} to the set of potentially incorrect answers \mathcal{P}_{t-1} to obtain the updated set \mathcal{P}_t .

Correction Phase During the correction phase, we use the set of potentially incorrect answers $\mathcal{P}_t = \{a_0, \dots, a_{t-1}\}$ as feedback to generate a corrected answer a_t . For a given question Q and the set \mathcal{P}_t , we append the phrase “the answer is likely not in \mathcal{P}_t ” to the question. This instructs the large language model to re-answer the question while avoiding repeating previous mistakes.

4.4 Final Answer Determination

The process of verify-then-correct can be iterated until specific stopping conditions are met. This process terminates under three situations: First, if the answer a_{t-1} is verified to be likely correct, it is selected as the final answer. Second, if the corrected answer a_t matches the previously generated answer a_{t-1} , then a_t is chosen as the final answer. Lastly, if the iteration count surpasses the maximum number of iterations T , the last LLM-generated answer a_T is adopted as the final answer.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate PROCO on three complex reasoning tasks: arithmetic reasoning (GSM8K (Cobbe et al., 2021b), AQuA (Ling et al., 2017), and MATH (Hendrycks et al., 2021)); open-domain question answering (NQ (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), WebQ (Berant et al., 2013), and HotpotQA (Yang et al., 2018)); and commonsense reasoning (CSQA (Talmor et al., 2019)). Detailed information about these datasets is available in Appendix A.1.

Baselines. We compare PROCO with three types of baselines: (1) LLM-generated documents: GenRead (Yu et al., 2023a). (2) Search engine-retrieved documents : RAG (Khattab et al., 2023). (3) Without external documents: CoT (Kojima et al., 2022), CoVe (Dhuliawala et al., 2023), and Self-Correct (Kim et al., 2023). All methods serve as baselines for open-domain question answering and commonsense reasoning tasks. For arithmetic reasoning, where external documents are unnecessary, CoT and Self-Correct are used. These baselines can be integrated into PROCO, for instance, using GenRead to generate an initial answer and PROCO to refine it (GenRead + PROCO). Details of all baselines are provided in Appendix A.2.

Evaluation Metrics. In open-domain question answering, we use exact match (EM) score and F1 score to evaluate model performance (Zhu et al., 2021). For other complex reasoning tasks, we use accuracy as the evaluation metric.

Implementation. We evaluate PROCO across three LLMs of different scales: GPT-3.5-Turbo-1106 and GPT-4-0125-Preview, which are the most widely used LLMs with public available APIs¹. Additionally, we include Mixtral-8x7B² (Jiang et al., 2024), an open source LLM with 47 billion parameters. For baselines like GenRead (Yu et al., 2023a) and RAG (Khattab et al., 2023) that use external documents, we set the number of documents $M = 5$. When incorporating these methods with PROCO, we set $M = 1$. The temperature parameter is set to 0.7 in our experiments.

5.2 Experimental Results

Overall performance on open-domain question answering and commonsense reasoning tasks.

¹<https://platform.openai.com/docs/models>

²<https://github.com/mistralai/mistral-src>

Method	Open-domain Question Answering								Commonsense Reasoning
	NQ		TriviaQA		WebQ		HotpotQA		CSQA
	EM	F1	EM	F1	EM	F1	EM	F1	Accuracy
<i>*Using LLMs to generate problem-related documents</i>									
GenRead	42.2 / 46.7	49.4 / 52.0	70.8 / 69.0	74.8 / 72.4	41.3 / 51.1	48.5 / 56.5	38.0 / 36.0	43.2 / 39.7	67.3 / 64.3
GenRead + PROCo	48.3 / 48.5	55.6 / 53.7	78.4 / 72.3	82.4 / 75.8	46.7 / 52.0	53.9 / 57.5	47.0 / 38.0	51.0 / 42.3	76.4 / 70.4
<i>*Using search engines to retrieve problem-related documents</i>									
RAG	45.3 / 48.8	52.4 / 54.6	72.7 / 75.3	76.4 / 78.5	40.1 / 46.3	46.9 / 52.1	37.0 / 37.0	41.1 / 40.2	65.9 / 66.3
RAG + PROCo	48.5 / 51.6	56.0 / 57.1	78.4 / 79.6	82.1 / 83.0	45.2 / 50.3	52.5 / 56.3	39.0 / 41.0	44.2 / 43.7	74.2 / 71.8
<i>*Direct question answering without external documents</i>									
CoT	40.3 / 42.6	46.4 / 48.2	69.2 / 66.7	72.2 / 70.3	38.2 / 46.6	44.6 / 51.9	28.0 / 29.0	31.2 / 34.4	72.9 / 68.4
Self-Correct	40.1 / 44.8	47.1 / 50.5	71.3 / 71.3	74.1 / 74.8	39.2 / 47.5	45.7 / 51.9	29.0 / 32.0	32.4 / 36.2	65.9 / 49.8
CoVe	43.4 / 47.6	48.9 / 53.0	76.4 / 73.2	79.4 / 76.4	43.1 / 53.4	49.0 / 58.2	31.0 / 33.0	35.2 / 36.9	73.1 / 70.8
PROCo	48.0 / 50.7	54.8 / 53.6	78.7 / 74.5	82.1 / 76.6	47.0 / 55.1	57.0 / 59.2	33.0 / 35.0	36.2 / 41.3	75.5 / 72.7

Table 2: Performance on NQ, TriviaQA, WebQ, HotpotQA, and CSQA benchmarks using GPT-3.5-Turbo-1106 (black-box LLM) and Mixtral-8x7B (open-source LLM). Each cell shows GPT-3.5-Turbo-1106 / Mixtral-8x7B performance. The best performance for each dataset is highlighted in bold. PROCo improves baseline methods with external documents across all benchmarks and outperforms those without external documents.

Method	Arithmetic Reasoning		
	GSM8K	AQuA	MATH
CoT	78.6 / 74.4	51.3 / 49.2	37.9 / 28.4
Self-Correct	75.1 / 72.5	48.7 / 44.4	27.6 / 21.5
PROCo	87.1 / 78.7	65.2 / 54.3	41.5 / 30.2

Table 3: Accuracy on arithmetic reasoning tasks. Each cell shows GPT-3.5-Turbo-1106 / Mixtral-8x7B performance. Since external documents are unnecessary for arithmetic reasoning, we only consider baseline methods without them. CoVe generates verification questions based on the semantics of the initial answer, which cannot be applied to numerical values.

Method	GSM8K	CSQA	HotpotQA
	Accuracy	Accuracy	EM
CoT	95.5	82.0	49.0
Self-Correct	91.5	79.5	49.0
CoVe	-	83.5	57.0
PROCo	97.6	86.7	61.0

Table 4: Performance comparison of various baseline methods using GPT-4-0125-Preview on three types of reasoning tasks: accuracy in GSM8K and CSQA, and EM score in HotpotQA.

Table 2 demonstrates that PROCo significantly enhances problem-solving performance across five benchmarks when combined with baseline methods using external documents. This improvement holds for both black-box and open-source LLM backends. Specifically, for GPT-3.5-Turbo-1106, using GenRead to generate an initial answer and then correcting it with PROCo (GenRead + PROCo) boosts the EM score by +6.1 on NQ, +7.6 on TriviaQA, +5.4 on WebQ, +9.0 on HotpotQA, and improves accuracy by +9.1 on CSQA.

Without external documents, PROCo shows superior self-correctness compared to Self-Correct and CoVe. It achieves gains of +7.9, +7.4, +7.8, +4.0, and +9.6 on NQ, TriviaQA, WebQ, HotpotQA, and CSQA, respectively, compared to Self-Correct. Additional experimental results are shown in Appendix A.5.

Overall performance on arithmetic reasoning tasks. For arithmetic reasoning tasks, we compare PROCo only with CoT and Self-Correct, as baselines with external documents and CoVe are unsuitable. As shown in Table 3, PROCo demonstrates superior self-correctness over all baseline methods across benchmarks on both black-box and open-source LLMs. Specifically, when applied to GPT-3.5-Turbo-1106, PROCo improves accuracy by an average of 14.1 compared to the Self-Correct.

PROCo with GPT-4 as backbone model. We compare PROCo with baseline methods using the GPT-4-0125-Preview model to test its effectiveness. Due to the high cost of GPT-4-0125-Preview, we select GSM8K for arithmetic reasoning, HotpotQA for open-domain question answering, and CSQA for commonsense reasoning. Only baseline methods without external documents are included. As shown in Table 4, PROCo outperforms the baselines across all benchmarks with the GPT-4 model.

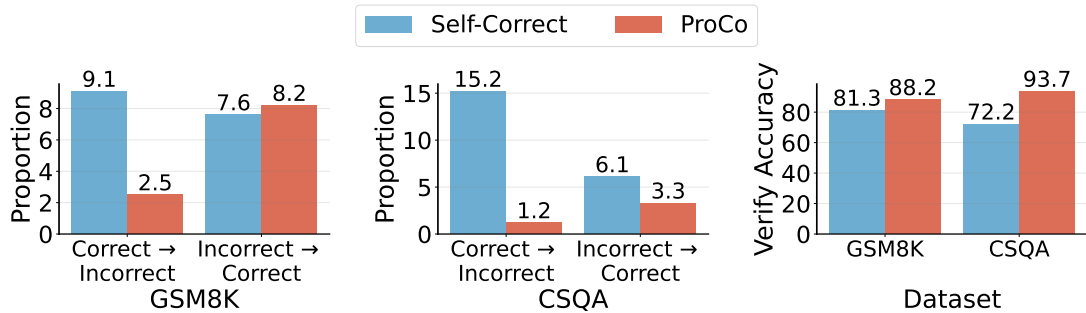


Figure 3: Analysis of answer changes after three correction rounds. Correct → Incorrect: A correct answer becomes incorrect. Incorrect → Correct: An incorrect answer is revised correctly. Self-Correct tends to change correct answers to incorrect ones rather than fixing errors. PROCO accurately judges and corrects wrong answers..

Method	NQ		TriviaQA		WebQ	
	EM	Tokens	EM	Tokens	EM	Tokens
GenRead	42.2	1023.3	70.8	924.2	41.3	963.3
GenRead + PROCO	48.3	469.1	78.4	465.0	46.7	416.8
Δ	14.5% ↑	54.2% ↓	10.7% ↑	49.7% ↓	13.1% ↑	56.7% ↓
RAG	45.3	1971.5	72.7	1937.5	40.1	2067.8
RAG + PROCO	48.5	916.4	78.4	968.2	45.2	875.5
Δ	7.1% ↑	53.5% ↓	7.8% ↑	50.0% ↓	12.7% ↑	57.7% ↓

Table 5: Comparison of PROCO with baselines including external documents: Efficiency and Effectiveness. PROCO consistently outperforms baselines on all benchmarks using significantly fewer tokens.

Retrieve External Documents vs. PROCO.

Since both retrieve external documents (RAG / GenRead) and verify-and-correct (PROCO) can enhance the performance on complex reasoning via adding the token cost for each question, we want to discuss the trade-off between efficiency and effectiveness to apply them to real-world reasoning task. Table 5 shows that PROCO outperforms GenRead/RAG in EM scores across three open-domain question-answering benchmarks, using just one external document compared to five. PROCO achieves an average 12.8% higher EM score than GenRead and 9.2% higher than RAG, while using half the tokens. Further analysis shows that multiple external documents often contain excessive irrelevant or redundant information, leading to incorrect answers and unnecessary token costs.

5.3 Analysis in PROCO

Analysis of Self-Correctness in PROCO Figure 3 shows the impact of PROCO after three correction rounds using GPT-3.5-Turbo-1106. PROCO is more accurate than Self-Correct in identifying errors in LLM-generated answers, with an 6.9% improvement in GSM8K and a 21.5% improvement in CSQA. Furthermore, PROCO is more effective at correcting errors without introducing new ones, enhancing LLM performance in complex reason-

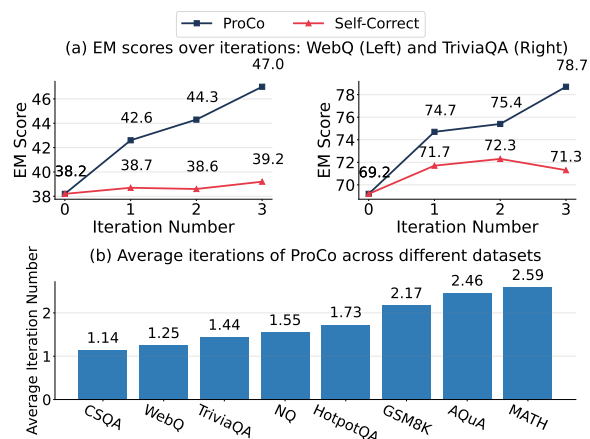


Figure 4: Analysis of Iterations in PROCO: PROCO shows improved performance with an increased number of iterations. Across all benchmarks, a minimum of three iterations ensures PROCO has ample opportunity to verify and correct answers generated by LLMs.

ing tasks. For GSM8K, PROCO incorrectly alters correct answers 2.5% of the time and corrects incorrect answers 8.2% of the time, compared to Self-Correct, which makes correct answers incorrect 9.1% of the time and fixes incorrect answers 7.6% of the time.

Impact of Iteration Count. Figure 4 (a) shows that the EM score of PROCO improves with more iterations in both WebQ and TriviaQA. In contrast, Self-Correct shows minimal improvement in WebQ and even a 1.0 EM score decrease in TriviaQA with more iterations. Additionally, the average iteration number for PROCO across all benchmarks is less than 3. Considering the trade-off between efficiency and effectiveness, we set the maximum number of iterations to 3 per question.

Comparison between Self-consistency and PROCO Since PROCO iteratively corrects an-

Question	Where does patience is a virtue come from?
Gold Answer	Psychomachia
CoT	The phrase “patience is a virtue” is an old proverb that has been passed down through generations. Its exact origin is not known, but it is believed to have originated from ancient wisdom and moral teachings . ✗
Self-Correct	The phrase “patience is a virtue” comes from the English poet William Langland’s work Piers Plowman in the 14th century. ✗
CoVe	“Patience is a virtue” is a well-known proverb that has been in use since at least the 14th century. It is often attributed to the Latin writer Publilius Syrus . ✗
PROCO	The phrase “patience is a virtue” can be traced back to medieval European literature. For instance, in the Psychomachia , a 5th-century poem by Prudentius, patience is depicted as one of the virtues. ✓
RAG	The first known recording of the expression in English is in the poem Piers Plowman by William Langland, written around the year 1360. Langland’s poem includes the line “Patience is a high virtue”. (https://www.bookbrowse.com/expressions/detail/index.cfm/expression_number/416/patience-is-a-virtue) ✗
RAG + PROCO	The idea behind the expression “patience is a virtue” dates back to the 5th century. It can be found in the epic poem Psychomachia . (https://writingexplained.org/idiom-dictionary/patience-is-a-virtue) ✓

Table 6: Case study of answers generated by various methods. The final answer is highlighted in yellow. PROCO shows superior self-correction compared to baseline methods that include self-correction processes. Additionally, PROCO reduces errors generated by methods that use external documents, ensuring correct source citation.

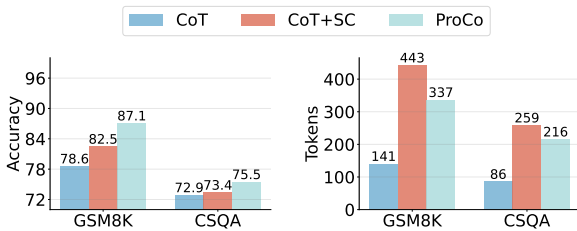


Figure 5: Performance comparison of CoT, PROCO, and CoT with self-consistency (CoT + SC). Compared to CoT + SC, PROCO not only exhibits higher accuracy but also consumes fewer tokens.

swers for complex reasoning tasks, we propose that Self-consistency (SC) (Wang et al., 2023c), which solves a problem multiple times and uses a majority vote to determine the final answer, may reduce errors by minimizing bias and enhancing the robustness of LLM performance.

We evaluate the performance of CoT with self-consistency (CoT + SC) on two complex reasoning tasks (GSM8K and CSQA) and compare it with PROCO. For a fair comparison, CoT + SC generates answers three times per question, matching ProCo’s maximum iterations. We find that PROCO uses fewer tokens and achieves better accuracy on both tasks. This is because, unlike PROCO’s verification and correctness processes, CoT + SC merely solves the problem multiple times, often repeating

the same mistakes.

5.4 Case Study

Table 6 shows that, except for RAG + PROCO and PROCO, all other methods fail to provide the correct answer to the given problem. CoT generates an incorrect answer, unable to determine the origin of the phrase “Patience is a virtue”. Self-Correct, CoVe, and RAG erroneously assert that the phrase originated in the 14th century. In contrast, RAG + PROCO and PROCO accurately identify the first appearance of the phrase “Patience is a virtue” in the 5th century. Furthermore, RAG + PROCO provides the correct source for citation. This indicates that integrating RAG into PROCO can significantly enhance the accuracy and reliability of answers.

6 Conclusion

In this study, we present a novel zero-shot prompting method for solving complex reasoning tasks. We name it progressive correction (PROCO), which first prompts an LLM to generate an initial response, then iterates a verify-then-correct process to progressively identify and correct (probably) false responses. Extensive experiments on eight complex reasoning datasets demonstrate the effectiveness and efficiency of our proposed method.

524
525
526
527
528
529
530

531

532
533
534
535
536
537

538
539
540
541

542
543
544
545
546

547
548
549
550
551
552

553
554
555
556
557
558

559
560
561
562

563
564
565
566
567
568

569
570
571
572
573

574
575

Limitations

This study focused exclusively on addressing complex reasoning tasks in English, with non-English tasks excluded from our training and test data. Consequently, the method may not perform well for non-English tasks. Future research will explore solutions for multilingual complex reasoning tasks.

References

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. [Semantic parsing on Freebase from question-answer pairs](#). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington. Association for Computational Linguistics.

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. 2024. [Teaching large language models to self-debug](#). In *The Twelfth International Conference on Learning Representations*.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. 2024. [Factool: Factuality detection in generative AI - a tool augmented framework for multi-task and multi-domain scenarios](#).

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021a. [Training verifiers to solve math word problems](#).

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

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. [Chain-of-verification reduces hallucination in large language models](#).

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. [SimCSE: Simple contrastive learning of sentence embeddings](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhibin Gou, Zhihong Shao, Yeyun Gong, yelong shen, Yujia Yang, Nan Duan, and Weizhu Chen. 2024. [CRITIC: Large language models can self-correct with tool-interactive critiquing](#). In *The Twelfth International Conference on Learning Representations*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. [Deberta: Decoding-enhanced](#)

[bert with disentangled attention](#). In *International Conference on Learning Representations*. 576
577

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. [Measuring mathematical problem solving with the math dataset](#). In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran. 578
579
580
581
582
583

Jie Huang and Kevin Chen-Chuan Chang. 2023. [Towards reasoning in large language models: A survey](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics. 584
585
586
587
588

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024. [Large language models cannot self-correct reasoning yet](#). In *The Twelfth International Conference on Learning Representations*. 589
590
591
592
593
594

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L elio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Th eophile Gervet, Thibaut Lavril, Thomas Wang, Timoth ee Lacroix, and William El Sayed. 2024. [Mixture of experts](#). 595
596
597
598
599
600
601
602
603
604
605

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. [TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics. 606
607
608
609
610
611
612

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2023. [Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp](#). 613
614
615
616
617

Geunwoo Kim, Pierre Baldi, and Stephen McAleer. 2023. [Language models can solve computer tasks](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 39648–39677. Curran Associates. 618
619
620
621

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. [Large language models are zero-shot reasoners](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc. 622
623
624
625
626

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob

632	Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research . <i>Transactions of the Association for Computational Linguistics</i> , 7:452–466.	689
633		690
634		691
635		692
636	Hung Le, Yue Wang, Akhilesh Deepak Gotmare, Silvio Savarese, and Steven Hoi. 2022. CodeRL: Mastering code generation through pretrained models and deep reinforcement learning . In <i>Advances in Neural Information Processing Systems</i> .	693
637		694
638		695
639		696
640		697
641	Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2023. Making language models better reasoners with step-aware verifier . In <i>Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 5315–5333, Toronto, Canada. Association for Computational Linguistics.	698
642		699
643		700
644		701
645		702
646		703
647		704
648	Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let’s verify step by step .	705
649		706
650		707
651		708
652	Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems . In <i>Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 158–167, Vancouver, Canada. Association for Computational Linguistics.	709
653		710
654		711
655		712
656		713
657		714
658		715
659	Jiacheng Liu, Skyler Hallinan, Ximing Lu, Pengfei He, Sean Welleck, Hannaneh Hajishirzi, and Yejin Choi. 2022. Rainier: Reinforced knowledge introspector for commonsense question answering . In <i>Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing</i> , pages 8938–8958, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.	716
660		717
661		718
662		719
663		720
664		721
665		722
666		723
667	Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback . In <i>Advances in Neural Information Processing Systems</i> , volume 36, pages 46534–46594. Curran Associates, Inc.	724
668		725
669		726
670		727
671		728
672		729
673		730
674		731
675		732
676	Ning Miao, Yee Whye Teh, and Tom Rainforth. 2024. Selfcheck: Using LLMs to zero-shot check their own step-by-step reasoning . In <i>The Twelfth International Conference on Learning Representations</i> .	733
677		734
678		735
679		736
680	Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback .	737
681		738
682		739
683		740
684		741
685		742
686	Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge . In <i>Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)</i> , pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.	743
687		744
688		745
	Niket Tandon, Aman Madaan, Peter Clark, Keisuke Sakaguchi, and Yiming Yang. 2021. Interscript: A dataset for interactive learning of scripts through error feedback .	
	Jinyuan Wang, Junlong Li, and Hai Zhao. 2023a. Self-prompted chain-of-thought on large language models for open-domain multi-hop reasoning . In <i>Findings of the Association for Computational Linguistics: EMNLP 2023</i> , pages 2717–2731, Singapore. Association for Computational Linguistics.	
	Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023b. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models . In <i>Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 2609–2634, Toronto, Canada. Association for Computational Linguistics.	
	Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023c. Self-consistency improves chain of thought reasoning in language models . In <i>The Eleventh International Conference on Learning Representations</i> .	
	Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models . In <i>Advances in Neural Information Processing Systems</i> , volume 35, pages 24824–24837. Curran Associates.	
	Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khoshabi, and Yejin Choi. 2023. Generating sequences by learning to self-correct . In <i>The Eleventh International Conference on Learning Representations</i> .	
	Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering . In <i>Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing</i> , pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.	
	Longhui Yu, Weisen Jiang, Han Shi, Jincheng YU, Zhengying Liu, Yu Zhang, James Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2024. Metamath: Bootstrap your own mathematical questions for large language models . In <i>The Twelfth International Conference on Learning Representations</i> .	
	Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu,	

746 Michael Zeng, and Meng Jiang. 2023a. [Generate](#)
747 [rather than retrieve: Large language models are](#)
748 [strong context generators](#). In *The Eleventh Inter-*
749 *national Conference on Learning Representations*.

750 Wenhao Yu, Zhihan Zhang, Zhenwen Liang, Meng
751 Jiang, and Ashish Sabharwal. 2023b. [Improving lan-](#)
752 [guage models via plug-and-play retrieval feedback](#).

753 Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex
754 Smola. 2023. [Automatic chain of thought prompting](#)
755 [in large language models](#). In *The Eleventh Interna-*
756 *tional Conference on Learning Representations*.

757 Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei,
758 Nathan Scales, Xuezhi Wang, Dale Schuurmans,
759 Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H.
760 Chi. 2023. [Least-to-most prompting enables com-](#)
761 [plex reasoning in large language models](#). In *The*
762 *Eleventh International Conference on Learning Rep-*
763 *resentations*.

764 Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming
765 Zheng, Soujanya Poria, and Tat-Seng Chua. 2021.
766 [Retrieving and reading: A comprehensive survey on](#)
767 [open-domain question answering](#).

A Appendix 768

A.1 Datasets 769

770 We evaluate PROCO on three complex reasoning
771 tasks: arithmetic reasoning (GSM8K (Cobbe et al.,
772 2021b), AQuA (Ling et al., 2017), and MATH
773 (Hendrycks et al., 2021)); open-domain question
774 answering (NQ (Kwiatkowski et al., 2019), Triv-
775 iaQA (Joshi et al., 2017), WebQ (Berant et al.,
776 2013), and HotpotQA (Yang et al., 2018)); and
777 commonsense reasoning (CSQA (Talmor et al.,
778 2019)). All of these datasets are accessible un-
779 der the MIT License. Below, we provide brief
780 descriptions of the datasets used:

- GSM8K (Cobbe et al., 2021b) consists of high
781 quality grade school math word problems cre-
782 ated by human problem writers. These prob-
783 lems require 2 to 8 steps to solve, and solu-
784 tions primarily involve performing a sequence
785 of elementary calculations using basic arith-
786 metic operations to reach the final answer. 787
- AQuA (Ling et al., 2017) contains multiple-
788 choice math questions that cover a broad
789 range of topics and difficulty levels. 790
- MATH (Hendrycks et al., 2021) is a challeng-
791 ing datasets consisting of 12k problems across
792 seven categories, testing models’ advanced
793 math and science reasoning. The problems in
794 this dataset are very hard as they come from
795 mathematics competitions written in \LaTeX . 796
- NQ (Kwiatkowski et al., 2019) were collected
797 from real Google search queries and the an-
798 swers are one or multiple spans in Wikipedia
799 articles identified by human annotators. 800
- TriviaQA (Joshi et al., 2017) includes trivia
801 questions with answers originally scraped
802 from trivia and quiz-league websites. 803
- WebQ (Berant et al., 2013) consists of ques-
804 tions selected using Google Suggest API,
805 where the answers are entities in Freebase. 806
- HotpotQA (Yang et al., 2018) contains 113k
807 multi-hop questions in natural language. The
808 questions are collected by crowdsourcing
809 based on Wikipedia articles with human anno-
810 tated supporting evidence and answers. 811
- CSQA (Talmor et al., 2019) offers a collection
812 of multiple-choice questions testing common-
813 sense reasoning. We use the development set
814 for our evaluation. 815

A.2 Baselines 816

817 To verify the effectiveness of our method, we com-
818 pare PROCO with three principal baseline cate-

gories:

- Using LLMs to generate problem-related documents: GenRead (Yu et al., 2023a) first prompts an LLM to generate M contextual documents based on a given question and then reads these documents to produce the final answer.
- Using search engines to retrieve problem-related documents: RAG (Khattab et al., 2023) first retrieves M relevant documents from Bing search³ based on a given question and then prompts an LLM to read the retrieved documents to produce the final answer.
- Direct question answering without external documents: CoT (Kojima et al., 2022) appends “Let’s think step by step” to the given question, instructing the LLM to generate a reasoning path leading to the final answer. CoVe (Dhuliawala et al., 2023) first answers the given question, generates a list of verification questions based on the initial answer, answers each of these verification questions, and finally produces the final answer based on the verification results. Self-Correct (Kim et al., 2023) instructs an LLM to critique and refine its initial response.

We use all methods as baselines for open-domain question answering and commonsense reasoning tasks. For arithmetic reasoning, where external documents are unnecessary, CoT and Self-Correct serve as baselines. These baseline methods can be integrated into PROCO. For example, we can use the GenRead (Yu et al., 2023a) method to generate an initial answer for a given question and use our proposed PROCO method to progressively correct the initial answer (i.e., GenRead + PROCO).

A.3 Evaluation Metrics

In open-domain question answering, we use exact match (EM) score and F1 score to evaluate model performance (Zhu et al., 2021). For the EM score, an answer is considered correct if and only if its normalized form (Yu et al., 2023a) has a match in the acceptable answer list. The F1 score treats the prediction and ground truth as bags of tokens, and computes the average overlap between them. For other complex reasoning tasks, we use accuracy as the evaluation metric.

³<https://www.microsoft.com/en-us/bing/apis/>

A.4 Full Prompts in Experiments

A.4.1 Arithmetic Reasoning

Given an arithmetic question Q , we use the CoT prompting method to generate an initial answer. Specifically, we first construct a reasoning generation prompt: “Q: Q . A: Let’s think step by step.” as shown in Prompt A.1. We then feed the above prompt to the LLM, which subsequently generates a reasoning path. To extract the answer from the reasoning path, we append an answer extraction instruction, creating the numerical answer extraction prompt: “Q: Q . A: {reasoning path} The answer (arabic numerals) is:” as shown in Prompt A.2.

Prompt A.1: Initial Answer Generation

Q: Q
A: Let’s think step by step.

Prompt A.2: Numerical Answer Extraction

Q: Q
A: {reasoning path} The answer (arabic numerals) is:

We use the substitute verification method to verify the correctness of the previous generated answer. Specifically, we first identify the key condition within the question (Sec. 3). By replacing the key condition with a specific token “X”, we create a masked question. We then append the sentence, “Suppose the answer is {previous generated answer}. What is the value of unknown variable X?” to the masked question to formulate the verification question, as shown in Prompt A.3.

Prompt A.3: Verification Question Construction

{masked question} Suppose the answer is {previous generated answer}. What is the value of unknown variable X?

Using Prompt A.1 and Prompt A.2, we can obtain the numerical answer for the verification question. By checking if the numerical answer for the verification question is equal to the key condition, we can assess the correctness of the previous generated answer. If the previous generated answer is deemed incorrect, we add it to the set of potentially incorrect answers; otherwise, we select it as the final answer. For incorrect answers, we can use the Prompt A.4 to correct them.

Prompt A.4: Incorrect Answers Correction

Q: Q (the answer is likely not in {set of potentially incorrect answers})
A: Let's think step by step.

A.4.2 Open-domain Question Answering

Given an open-domain question Q , we use the Prompt A.2 to instruct the LLM to generate a reasoning path. To extract the answer from this reasoning path, we add an answer extraction instruction, resulting in the following entity answer extraction prompt: "Answer the following question with just one entity. Q: Q . A: {reasoning path} The answer is:" as shown in Prompt A.5.

Prompt A.5: Initial Answer Generation

Answer the following question with just one entity.
Q: Q
A: {reasoning path} The answer is:

We use the substitute verification method to verify the correctness of the previous generated answer. Specifically, we first use the Prompt A.6 to identify the key condition within the question. By replacing the key condition with a specific token X , we create a masked question. We then append the sentence, "Suppose the answer is {previous generated answer}. What is the value of unknown variable X ?" to the masked question to formulate the verification question, as shown in Prompt A.3.

Prompt A.6: Key Condition Identification

Given the question below, the task is to identify a set of entities within the question and then select the one that is most relevant to the problem-solving process.
 Q

Using Prompt A.1 and Prompt A.5, we can obtain the answer for the verification question. By checking if the answer for the verification question and the key condition are equivalent, we can assess the correctness of the previous generated answer.

Prompt A.7: Equivalence Check

Determine the correctness of the proposition: If the answer to question {verification question} is {key condition}, then X could also be {answer for the verification question}

If the previous generated answer is deemed incorrect, we add it to the set of potentially incorrect answers; otherwise, we select it as the final answer. For incorrect answers, we can use the Prompt A.4 to correct them.

A.5 Additional Experimental Results

Can we just use the exact match method during the verification phase?

Since verification questions can have multiple valid answers, directly checking if the LLM-generated response exactly matches the key condition might misclassify correct answers as incorrect. Consider the following example: Given an open-domain question "Who wrote the treasure of the sierra madre?", we first prompt an LLM to generate an initial answer, e.g., "B. Traven". Next, we identify a key condition in the question relevant to the problem-solving process, such as "the treasure of the sierra madre". By masking the key condition, we create a verification question: "Who wrote X ? Suppose the answer is B. Traven. What is the value of unknown variable X ?". Using the LLM to solve the verification question, we receive the response "The Death Ship". If we directly check whether "The Death Ship" matches "the treasure of the sierra madre", we find they do not match, leading us to incorrectly judge the answer "B. Traven" as wrong. However, all books written by B. Traven are correct answers to the verification question. Thus, exact matching is insufficient for verification. Based on this observation, we propose proposition-based verification. Specifically, we construct an answer verification prompt: "Determine the correctness of the proposition: If the answer to question "Who wrote X ? Suppose the answer is B. Traven. What is the value of unknown variable X ?" is "the treasure of the sierra madre", then X could also be "The Death Ship"". We input this prompt into an LLM and receive a judgement about the proposition's correctness, e.g., "The proposition is correct, since both works were written by the same author.". This approach allows the LLM to properly analyze whether "The Death Ship" and "the treasure of the sierra madre" are both correct answers for the verification question, thus accurately determining the correctness of LLM-generated answers.

A.6 Sample Predictions for Complex Reasoning Datasets

In this section, we provide sample predictions generated by the PROCO prompting method. For readability, we adjusted line breaks in the original text. As shown in Tables 7 to 14, PROCO accurately identifies and rectifies errors in LLM-generated answers without external feedback. Additionally, integrating RAG into PROCO significantly improves the accuracy and reliability of the answers.

Question	Jean is two years older than Mark. Two years ago Mark was 5 years older than half Jan's age. If Jan is 30 how old is Jean?
Gold Answer	$\frac{30-2}{2} + 5 + 2 + 2 = 23$
Key Condition	5
Initial Answer	Two years ago, Mark was 5 years older than half of Jan's age. Half of Jan's age two years ago would be 10 ($30/2 - 2$). Therefore, Mark's age two years ago would be 15 ($10 + 5$). We know that Jean is two years older than Mark. Therefore, Jean's age two years ago would be 17 ($15 + 2$). Jean's current age would be 19 ($17 + 2$). ✗
Verification Question	Jean is two years older than Mark. Two years ago Mark was X years older than half Jan's age. If Jan is 30 how old is Jean? Suppose the answer is 19 . What is the value of unknown variable X?
Answer to the Verification Question	Two years ago Mark's age = Jean's age - 2 Half of Jan's age = 15 Mark's age two years ago = Jean's age - 2 = 15 - X Therefore, X = Jean's age - 15 - 2 = 19 - 15 - 2 = 2. Answer: X = 2 .
Verification Result	Since 2 is not equal to the key condition, the answer 19 is likely incorrect .
Question (with Hint)	Jean is two years older than Mark. Two years ago Mark was 5 years older than half Jan's age. If Jan is 30 how old is Jean? (The answer is likely not 19)
Corrected Answer	Two years ago, Mark was 5 years older than half of Jan's age. Jan was 28. Half of Jan's age two years ago would be 14. Mark was 5 years older than 14 two years ago, so he was 19 two years ago. Jean is two years older than Mark, so Jean was 21 two years ago. Now, two years have passed since then, so Jean is 23 today. ✓
Verification Question	Jean is two years older than Mark. Two years ago Mark was X years older than half Jan's age. If Jan is 30 how old is Jean? Suppose the answer is 23 . What is the value of unknown variable X?
Answer to the Verification Question	First, we're given that Jean is two years older than Mark. So if Mark is M years old, Jean is $M + 2$ years old. Two years ago, Mark was $M - 2$ years old and Jan's age was $30 - 2 = 28$ years. We're also given that two years ago Mark was X years older than half of Jan's age. So, $M - 2 = X + \frac{1}{2}(28)$. Now, let's solve for M using the information given. $M - 2 = X + 14$ $M = X + 16$ Now we know Mark's age in terms of X . Since Jean is two years older than Mark, Jean's age is $X + 18$. If the answer is 23 for Jean's age, then: $X + 18 = 23$ $X = 5$ So, the value of the unknown variable X is 5 .
Verification Result	Since 5 is equal to the key condition, the answer 23 is likely correct .
Final Answer	23 ✓

Table 7: Example output obtained by PROCO on the GSM8K dataset.

Question	In a class of 50 students, 28 participate in MATHCOUNTS, 21 participate in science club, and 6 students participate in neither. How many students participate in both MATHCOUNTS and science club?
Gold Answer	5
Key Condition	28
Initial Answer	Step 1: Find the total number of students who participate in either MATHCOUNTS or science club. Total students participating = $28 + 21 - 6$ Total students participating = 43 Step 2: Calculate the number of students who participate in both MATHCOUNTS and science club. Number of students in both = $28 + 21 - 43$ Number of students in both = 6 So, 6 students participate in both MATHCOUNTS and science club. ✗
Verification Question	First, we can calculate the number of students participating in either MATHCOUNTS or the science club or both: $50 - 6 = 44$. Next, we use the principle of inclusion and exclusion to set up our equation. The principle of inclusion and exclusion for two sets states: $ A \cup B = A + B - A \cap B $ Here, $ A \cup B $ represents the total number of students participating in either MATHCOUNTS or science club or both, which we have found to be 44. $ A $ represents the number of students in MATHCOUNTS (X), $ B $ represents the number of students in science club (21), and $ A \cap B $ represents the number of students in both clubs (6). Substituting these values into the equation, we get: $44 = X + 21 - 6$. Thus, the value of X is 29 .
Verification Result	Since 29 is not equal to the key condition, the answer 6 is likely incorrect.
Question (with Hint)	In a class of 50 students, 28 participate in MATHCOUNTS, 21 participate in science club, and 6 students participate in neither. How many students participate in both MATHCOUNTS and science club? (The answer is likely not 6)
Corrected Answer	- Let M represent the set of students participating in MATHCOUNTS. - Let S represent the set of students participating in the science club. - Let n represent the total number of students in the class, which is 50. First, we calculate the number of students participating in either MATHCOUNTS or the science club or both: $ M \cup S = n - N = 50 - 6 = 44$ Using the principle of inclusion-exclusion for the union of two sets: $ M \cup S = M + S - M \cap S $ Substitute the known values into the equation: $44 = 28 + 21 - M \cap S $ $ M \cap S = 5$ Thus, the number of students who participate in both MATHCOUNTS and the science club is: 5 ✓

Table 8: Example output obtained by PROCO on the MATH dataset.

Question	A man spends Rs. 3500 per month and saves 12 1/2% of his income. His monthly income is ? Answer Choices: (A) Rs. 4400 (B) Rs. 4270 (C) Rs. 4000 (D) Rs. 3937.50 (E) None of these
Gold Answer	C
Key Condition	3500
Initial Answer	First, we need to calculate the amount saved each month. To do this, we can calculate 12 1/2% of Rs. 3500. $12\frac{1}{2}\%$ of Rs. 3500 = $(12.5/100) * 3500 = 0.125 * 3500 =$ Rs. 437.50 So, the man saves Rs. 437.50 each month. To find his monthly income, we can use the formula: Income = Expenditure + Savings Income = Rs. 3500 + Rs. 437.50 Income = Rs. 3937.50 Therefore, his monthly income is Rs. 3937.50, which corresponds to answer choice (D) . \times
Verification Question	A man spends Rs. X per month and saves 12 1/2% of his income. His monthly income is ? Answer Choices: (A) Rs. 4400 (B) Rs. 4270 (C) Rs. 4000 (D) Rs. 3937.50 (E) None of these Suppose the answer is (D) Rs. 3937.50. What is the value of unknown variable X?
Answer to the Verification Question	<p>1. Convert the percentage saved into a decimal: 12.5% can be expressed as:</p> $12.5\% = \frac{12.5}{100} = 0.125$ <p>2. Determine the amount saved: Since he saves 12.5% of his income, the amount saved each month is:</p> $0.125 \times 3937.50 = 492.1875$ <p>3. Calculate the amount spent (X): The amount he spends is the remaining part of his income after saving. Therefore, the amount spent X is:</p> $X = \text{Income} - \text{Amount saved} = 3445.3125$ <p>The value of the unknown variable X (the amount spent per month) is: 3445.3125</p>
Verification Result	Since 3445.3125 is not equal to the key condition, the answer (D) Rs. 3937.50 is likely incorrect.
Question (with Hint)	A man spends Rs. 3500 per month and saves 12 1/2% of his income. His monthly income is ? Answer Choices: (A) Rs. 4400 (B) Rs. 4270 (C) Rs. 4000 (D) Rs. 3937.50 (E) None of these (The answer is likely not (D) Rs. 3937.50)
Corrected Answer	<p>Let I be the man's monthly income. Since he saves 12.5% of his income, he spends the remaining percentage of his income. The percentage of income spent is:</p> $100\% - 12.5\% = 87.5\%$ $\text{Income spent} = 0.875 \times I$ <p>We know that the amount spent is Rs. 3500. Thus, we set up the equation:</p> $0.875 \times I = 3500$ <p>To find I, we solve the equation for I:</p> $I = \frac{3500}{0.875} = 4000$ <p>The man's monthly income is Rs. 4000. The correct answer is: (C) Rs. 4000 \checkmark</p>

Table 9: Example output obtained by PROCO on the AQuA dataset.

Question	Right to property according to the constitution of india is a?
Gold Answer	constitutional right
Key Condition	property
Initial Answer	The right to property in the Constitution of India has undergone significant changes over the years. Prior to the 44th Amendment Act in 1978, the right to property was considered a fundamental right under Part III of the Constitution. However, this amendment removed the right to property from the list of Fundamental Rights and converted it into a simple legal right (https://byjus.com/question-answer/consider-the-following-statements-regarding-right-to-property-1-right-to-property-is-legal-right/) ✗
Verification Question	Right to X according to the constitution of india is a? Suppose the answer is simple legal right. What is the value of unknown variable X?
Answer to the Verification Question	the right to legal representation.
Verification Result	Since the right to legal representation and the property are not equivalent, the answer simple legal right is likely incorrect.
Question (with Hint)	Right to property according to the constitution of india is a? (The answer is likely not simple legal right)
Corrected Answer	The right to property in the Constitution of India has evolved from being a fundamental right to a legal and constitutional right , with provisions for legal protection and acquisition for public welfare. (https://testbook.com/question-answer/what-is-the-position-of-the-right-to-property-in-i-616294b1fb5e57239764c589) ✓
Verification Question	Right to X according to the constitution of india is a? Suppose the answer is constitutional right. What is the value of unknown variable X?
Answer to the Verification Question	In the context of the Constitution of India, the unknown variable X refers to a specific right that is classified as a "constitutional right." Given this information, one possible value for X is "property." To elaborate, the right to property was originally a fundamental right under Article 31 of the Indian Constitution. However, it was later amended by the 44th Amendment Act of 1978, and the right to property was removed from the list of fundamental rights and reclassified as a constitutional right under Article 300A. Therefore, the value of the unknown variable X is "property."
Verification Result	Since the property and the property are equivalent, the answer constitutional right is likely correct.
Final Answer	constitutional right ✓

Table 10: Example output obtained by RAG + PROCO on the NQ dataset.

Question	Who was the defending champion when Martina Navratilova first won Wimbledon singles?
Gold Answer	Virginia Wade
Key Condition	Martina Navratilova
Initial Answer	When Martina Navratilova first won the Wimbledon singles title, the defending champion was Chris Evert (https://en.wikipedia.org/wiki/1977_Wimbledon_Championships). ✗
Verification Question	Who was the defending champion when X first won Wimbledon singles? Suppose the answer is Chris Evert. What is the value of unknown variable X?
Answer to the Verification Question	Given that Chris Evert is the answer, X must be the name of a person who won Wimbledon singles. Conclusion: X refers to a person who won Wimbledon singles, but without more information, the specific name cannot be determined.
Verification Result	Since unknown and Martina Navratilova are not equivalent, the answer Chris Evert is likely incorrect.
Question (with Hint)	Who was the defending champion when Martina Navratilova first won Wimbledon singles? (The answer is likely not Chris Evert)
Corrected Answer	The defending champion when Martina Navratilova first won the Wimbledon singles title in 1978 was Virginia Wade (https://en.wikipedia.org/wiki/1978_Wimbledon_Championships_-_Women's_singles). ✓
Verification Question	Who was the defending champion when X first won Wimbledon singles? Suppose the answer is Virginia Wade. What is the value of unknown variable X?
Answer to the Verification Question	Martina Navratilova won her first Wimbledon singles title in 1978. Virginia Wade was the Wimbledon singles champion in 1977. So, if we are to match the scenario where X first won Wimbledon when Virginia Wade was the defending champion: In 1978, the defending champion was Virginia Wade (from her 1977 win). Martina Navratilova won Wimbledon singles in 1978. Therefore, (X) is indeed Martina Navratilova. Thus, the value of the unknown variable (X) is Martina Navratilova.
Verification Result	Since Martina Navratilova and Martina Navratilova are equivalent, the answer Virginia Wade is likely correct.
Final Answer	Virginia Wade ✓

Table 11: Example output obtained by RAG + PROCO on the TriviaQA dataset.

Question	What highschool did harper lee go to?
Gold Answer	Monroe County High School
Key Condition	highschool
Initial Answer	Harper Lee attended several educational institutions during her academic journey. According to the information available, she attended Huntingdon College from 1944 to 1945. After that, she studied law at the University of Alabama from 1945 to 1949. Additionally, she spent a year at Oxford University (https://www.tokillamockingbird.com/harper-lee). Based on the available information, it can be concluded that Harper Lee did not attend high school. She pursued higher education after completing her grade school education. Conclusion Harper Lee did not attend high school. ✗
Verification Question	What X did harper lee go to? Suppose the answer is Harper Lee did not attend high school. What is the value of unknown variable X?
Answer to the Verification Question	If the statement "Harper Lee did not attend high school" is true, then the value of the unknown variable X, representing the high school she attended, would be "none." This means Harper Lee did not go to any high school.
Verification Result	Since none and highschool are not equivalent, the answer Harper Lee did not attend high school is likely incorrect.
Question (with Hint)	What highschool did harper lee go to? (The answer is likely not Harper Lee did not attend high school)
Corrected Answer	Nelle Harper Lee, the acclaimed author of "To Kill a Mockingbird," attended monroe county high school in her hometown of Monroeville, Alabama. She was born on April 28, 1926, and grew up in a rural southern Alabama town, where she attended both elementary school and high school. Her father, Amasa Coleman Lee, was an attorney who served in the state legislature in Alabama, and her childhood experiences were similar to those of her young narrator in "To Kill a Mockingbird," Scout Finch (https://www.chipublib.org/harper-lee-biography/) ✓
Verification Question	What X did harper lee go to? Suppose the answer is monroe county high school. What is the value of unknown variable X?
Answer to the Verification Question	If the answer is "Monroe County High School," the value of the unknown variable X is "high school."
Verification Result	Since high school and highschool are equivalent, the answer monroe county high school is likely correct.
Final Answer	monroe county high school ✓

Table 12: Example output obtained by RAG + PROCO on the WebQ dataset.

Question	What could go on top of wood? Answer Choices: (A) lumberyard, (B) synagogue, (C) floor, (D) carpet, (E) hardware store
Gold Answer	D
Key Condition	wood
Initial Answer	To determine what could go on top of wood, let's analyze the answer choices provided: (A) Lumberyard: A lumberyard is a facility where wood and lumber are stored and sold. It is not something that goes on top of wood. (B) Synagogue: A synagogue is a place of worship for Jewish people. It is not something that goes on top of wood. (C) Floor: Yes, a floor can go on top of wood. Wood is commonly used as a material for flooring, providing a natural and warm aesthetic. Hardwood floors, for example, are made of solid wood and are installed on top of a subfloor. (D) Carpet: Yes, carpet can go on top of wood. Carpet is often installed on top of a wooden subfloor, providing a soft and comfortable surface. (E) Hardware store: A hardware store is a retail store that sells tools, building materials, and other hardware items. It is not something that goes on top of wood. Therefore, the correct answer is (C) floor . ✗
Verification Question	What could go on top of X? Suppose the answer is floor. What is the value of unknown variable X?
Answer to the Verification Question	The question asks what could go on top of X, and the answer is "floor." Based on this, we can infer that X refers to a "ceiling." The ceiling is the upper surface of a room and is what goes on top of the floor. Therefore, X refers to ceiling .
Verification Result	Since wood and ceiling are not equivalent, the answer (C) floor is likely incorrect .
Question (with Hint)	What could go on top of wood? Answer Choices: (A) lumberyard, (B) synagogue, (C) floor, (D) carpet, (E) hardware store (The answer is likely not (C) floor)
Corrected Answer	The answer is likely not in ['C']. Based on the provided search results, the item that could go on top of wood is (D) carpet . ✓
Verification Question	What could go on top of X? Suppose the answer is carpet. What is the value of unknown variable X?
Answer to the Verification Question	If the answer to "What could go on top of X?" is "carpet," then X is likely the floor or flooring. Therefore, the value of the unknown variable X is floor .
Verification Result	Yes, the proposition is correct. If the answer to the question "What could go on top of X? Suppose the answer is carpet" is "wood", then X could also be "floor". Here's the reasoning: 1. The question asks what could go on top of X, and the given answer is "carpet". 2. This implies that "carpet" can be placed on top of X. 3. The follow-up statement tells us that the answer to this is "wood". Therefore, wood is X. 4. Carpet is typically placed on top of a floor surface. 5. Wood is a common type of flooring. Thus, if wood is the answer to X, and considering the context of flooring, X being "floor" is a reasonable interpretation because a carpet is often placed on a wood floor. Therefore, the unknown variable X, which in this context is the surface on which carpet can be placed, could be "floor". Since floor and wood are equivalent, the answer (D) carpet is likely correct .
Final Answer	D ✓

Table 13: Example output obtained by PROCO on the CSQA dataset.

Question	The trophy given to the winner of the college football game between University of Idaho Vandals and University of Montana Grizzlies is derived from which European culture?
Gold Answer	German
Key Condition	European culture
Initial Answer	The University of Idaho Vandals and the University of Montana Grizzlies are both located in the United States. However, the trophy's origin may be derived from a European culture due to the historical influence of European traditions on American sports and collegiate traditions. ✗
Verification Question	The trophy given to the winner of the college football game between University of Idaho Vandals and University of Montana Grizzlies is derived from which X ? Suppose the answer is European culture. What is the value of unknown variable X?
Answer to the Verification Question	Determine which aspect of European culture matches the question about the trophy. Given that the trophy is derived from European culture, X must be a specific aspect or tradition within European culture. X is " Medieval armor and weaponry ".
Verification Result	Since Medieval armor and weaponry and European culture are not equivalent, the answer European culture is likely incorrect.
Question (with Hint)	The trophy given to the winner of the college football game between University of Idaho Vandals and University of Montana Grizzlies is derived from which European culture? (The answer is likely not European culture)
Corrected Answer	First, it's important to note that the University of Idaho Vandals and the University of Montana Grizzlies are both located in the United States, so the trophy is likely derived from American culture rather than European culture. The trophy for this game is called the Little Brown Stein, and it is a tradition that dates back to 1938. The name "Stein" suggests a German influence, as "stein" is a German word for "stone" or "mug." Therefore, the trophy is likely derived from German culture. ✓
Verification Question	The trophy given to the winner of the college football game between University of Idaho Vandals and University of Montana Grizzlies is derived from which X ? Suppose the answer is German culture. What is the value of unknown variable X?
Answer to the Verification Question	If the trophy given to the winner of the college football game between the University of Idaho Vandals and the University of Montana Grizzlies is derived from German culture, the unknown variable X could be " German traditions. "
Verification Result	Given the information provided, if the answer to the question is indeed "European culture", then X could indeed be "German traditions". This is because German traditions could fall under the umbrella of European culture. German traditions are a subset of European culture, so if the answer is broadened from "European culture" to a more specific subset like "German traditions", it would still be consistent with the broader category. Therefore, the proposition appears to be correct. Since European culture and German traditions are equivalent, the answer German culture is likely correct.
Final Answer	German culture ✓

Table 14: Example output obtained by PROCO on the HotpotQA dataset.