

Error Reflection Prompting: Can Large Language Models Successfully Understand Errors?

Jason Li Lauren Yraola Kevin Zhu Sean O'Brien

Algoverse AI Research

j1255788@gmail.com, kevin@algoverse.us

Abstract

Prompting methods for language models, such as Chain-of-thought (CoT), present intuitive step-by-step processes for problem solving. These methodologies aim to equip models with a better understanding of the correct procedures for addressing a given task. Despite these advancements, CoT lacks the ability of reflection and error correction, potentially causing a model to perpetuate mistakes and errors. Therefore, inspired by the human ability for said tasks, we propose Error Reflection Prompting (ERP) to further enhance reasoning in language models. Building upon CoT, ERP is a method comprised of an incorrect answer, error recognition, and a correct answer. This process enables the model to recognize types of errors and the steps that lead to incorrect answers, allowing the model to better discern which steps to avoid and which to take. The model is able to generate the error outlines itself with automated ERP generation, allowing for error recognition and correction to be integrated into the reasoning chain and produce scalability and reliability in the process. The results demonstrate that ERP serves as a versatile supplement to conventional CoT, ultimately contributing to more robust and capable reasoning abilities along with increased interpretability in how models ultimately reach their errors.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable performance across numerous NLP tasks (OpenAI, 2023). Yet scaling them is costly, prompting interest in more efficient methods like Chain-of-Thought (CoT) (Wei et al., 2023), which guides step-by-step reasoning. A key shortcoming of CoT, however, is that it lacks a mechanism to recognize and correct its own mistakes (Turpin et al., 2023). This can lead models to persist with errors, as they lack explicit guidance to reflect on and revise flawed reasoning.

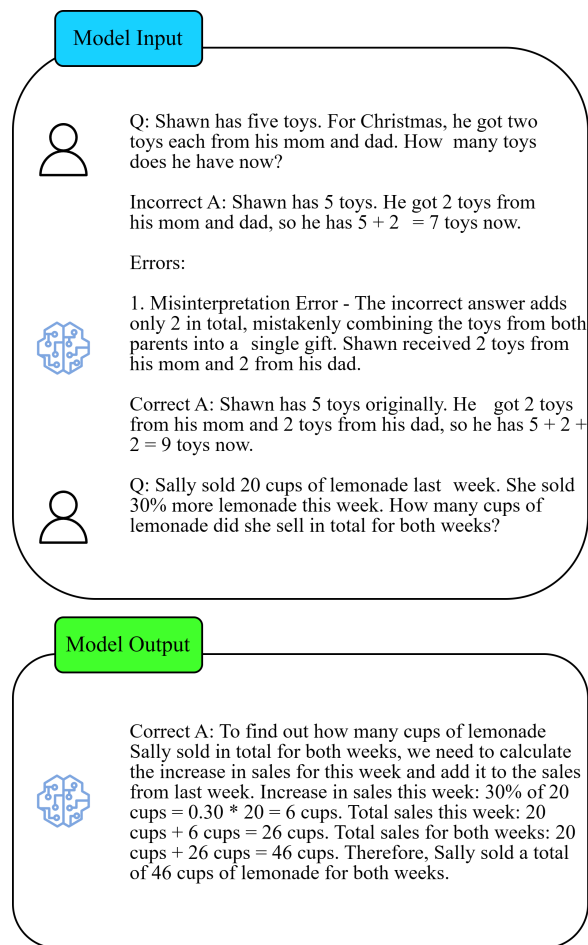


Figure 1: Example of Error Reflection Prompting with a focus on misinterpretation.

Drawing on the human ability to self-reflect and amend mistakes (Huang and Chang, 2023), we propose Error Reflection Prompting (ERP). ERP supplies an incorrect answer alongside an explanation of the errors, then presents the correct reasoning. By exposing common pitfalls and clarifying why they are wrong, ERP steers models away from these errors in future reasoning. Evaluations on multiple benchmarks with GPT-3.5 and GPT-4 confirm that ERP consistently outperforms standard CoT, notably lowering specific error types (e.g., algebraic

and calculation mistakes) and balancing the overall error distribution. Thus, ERP adds a complementary layer of self-correction to CoT, enhancing both accuracy and transparency in model outputs.

Dataset	Type of Reasoning	Test
GSM8K	Arithmetic	1319
AQuA	Arithmetic	254
MATH	Arithmetic	1000
CSQA	Commonsense	1221
StrategyQA	Commonsense	2290

Table 1: Problems used from each dataset. Note that CSQA and StrategyQA’s test set answers are not publicly available, so we follow (Wei et al., 2023) and evaluate performance on development (dev) datasets.

We propose **Error Reflection Prompting (ERP)**, which prompts an LLM with an incorrect answer, explicit errors, and a corrected solution. This strategy helps the model identify and avoid known pitfalls:

1. **Incorrect Answer.** Provide a solution with specific errors—ones the model commonly makes (Gao et al., 2022)—ensuring the prompt focuses on these weaknesses.
2. **Error Reflection.** Explain each error and why it is incorrect, offering rationales to help the model understand and avoid repeating them.
3. **Correct Answer.** Present the corrected chain of reasoning, reinforcing the proper approach.

Formally, let E be a set of errors identified during analysis, and $Q \subseteq E$ be a subset of n errors. For each example, we embed Q into the prompt and provide corresponding explanations and the correct solution.

In contrast to a single, general CoT prompt, ERP may need unique prompts for each problem, which can be time-consuming. To make ERP scalable, we automate the process:

1. Prompt the model to generate n plausible errors for a given question.
2. Construct an incorrect solution incorporating those errors.
3. Incorporate the question, incorrect solution, errors, and correct solution into the final prompt.

By letting the model generate its own error outlines, ERP integrates recognition and correction into the reasoning chain, making it more practical for large-scale tasks.

2 Experiments

2.1 Experimental Setup

We focus our analysis on two types of reasoning: commonsense and arithmetic. By using random sampling, we ran tests on all shuffled datasets.

Dataset details are included in Table 1.

Arithmetic Dataset For arithmetic reasoning, we consider three of the following datasets: GSM8K, a dataset for grade school math word problems (Cobbe et al., 2021), AQuA, a dataset consisting of algebraic word problems (Ling et al., 2017), and MATH, a dataset with challenging competition math problems (Hendrycks et al., 2021). All datasets have been run on their respective test sets except for MATH, for which we have limited to 1000 test samples for a reasonable computing limit.

Commonsense Datasets In regards to commonsense reasoning, we conduct tests on StrategyQA, which requires models to infer solutions to questions with implicit reasoning steps (Geva et al., 2021), and CSQA, a dataset for commonsense question answering (Talmor et al., 2019).

Models We use the OpenAI Chat Completions API, one of the most widely known and used frameworks for accessing language models (Ge et al., 2023). We use GPT-3.5 and GPT-4 for all tests. Specifically, we use gpt-3.5-turbo-0613 and gpt-4-1106-preview respectively for all tests. In addition, we employ 4-shot prompts for each dataset, and each prompt is derived from previous works in (Wei et al., 2023). All errors were created from human annotators or generated using GPT-4 with some slight modifications. Full prompts can be found in Appendix A.

2.2 Analysis Evaluation

After collecting the results, we analyze the distribution of errors according to different categories. To accomplish our analysis of the model’s errors, GPT-4 was used as an annotator. A sample of 100 errors from each dataset was pulled to evaluate the effectiveness of GPT-4 as an error detector. To categorize each error and understand the underlying distribution, we separated errors into five and four different classes for arithmetic and com-

Prompting Method	Arithmetic Reasoning		
	GSM8K Accuracy	AQuA Accuracy	MATH Accuracy
GPT-3.5 with 5-shot Auto ERP	79.8%	48.0%	—
GPT-3.5 with 4-shot CoT	74.6%	54.3%	40.2%
GPT-3.5 with 4-shot ERP	77.8%	58.7%	41.1%
GPT-4 with 4-shot CoT	95.4%	75.9%	54.1%
GPT-4 with 4-shot ERP	95.7%	76.7%	54.8%

Table 2: Results from a variety of math-based datasets

Prompting Method	Commonsense Reasoning	
	StrategyQA Acc.	CSQA Acc.
GPT-3.5 with 4-shot CoT	65.3%	75.8%
GPT-3.5 with 4-shot ERP	66.1%	77.3%
GPT-3.5 with 4-shot ERP w/ Categories	62.9%	—
GPT-4 with 4-shot CoT	79.9%	85.3%
GPT-4 with 4-shot ERP	80.0%	86.3%
GPT-4 with 4-shot ERP w/ Categories	76.0%	—

Table 3: Results from commonsense-based datasets

monsense reasoning datasets respectively. For the arithmetic datasets, we have **Misinterpretation of Question, Missed Steps, Logical/Commonsense Error, Wrong Selection of Answer, and Algebraic/Calculation Error**. For more details on each of these categories, please see Table B in the Appendix. As for the commonsense datasets, we have taken a majority of our errors from the categories presented in (Dou et al., 2022) being **Commonsense Error, Encyclopedic Error, and Self-Contradiction Error**. We have also added an **Assumption Error** to better align with problem solving tasks. We have chosen to analyze on all datasets except for the MATH dataset (Hendrycks et al., 2021) as the model may struggle to identify errors due to the complexity of the problems. All of these errors and their examples are displayed in Table 4.

3 Results

3.1 Arithmetic Results

ERP demonstrates increases over conventional CoT. Table 2 shows the differences between the two prompting methods. The largest gain is in AQuA, outperforming CoT by **+4.4%**. However, with the automatic prompt generation, ERP is able to outperform CoT by **+5.2%**. Though ERP outperforms CoT on all three datasets, the MATH (Hendrycks et al., 2021) dataset had the lowest per-

centage increase. It is the most difficult dataset in terms of problem solving ability, and a strong reasoning ability is required to score well on it.

The correlation between error complexity and problem difficulty should be further researched to understand whether or not ERP may be effective with more complicated errors. Error outlining may not have been effective at reflection due to the discrepancy in difficulty and complexity.

3.2 Commonsense Results

ERP scores higher on commonsense data versus CoT. GPT-4 + ERP had a slight increase over conventional CoT in CSQA, outperforming it by **+1.0%** while GPT-3.5 + ERP had an increase of **+1.5%**. For commonsense reasoning, ERP appears less effective than arithmetic reasoning. However, experimentation with the types of errors presented in the prompt may aid in reasoning.

4 Error Analysis

In this section, we explore how ERP affects the distribution of errors in answering questions. Although ERP scored higher than CoT (Wei et al., 2023) for all the tests, the error distribution was not the same. This discrepancy is analyzed to explain ERP’s behavior in a LLM. Confidence values for annotations were calculated by taking a subset of fifty randomly sampled annotated errors from both GPT-3.5 and GPT-4 inference errors and determining their validity. These values can be found in Table 5. Let $T = \text{GPT-3.5 errors}$ and let $F = \text{GPT-4 errors}$ where both $T = \{0, 1\}$ and $F = \{0, 1\}$:

$$\frac{1}{100} \left(\sum_{i=1}^{50} T_i + \sum_{i=1}^{50} F_i \right)$$

4.1 Arithmetic Errors

Figure 2 shows a set of errors from the annotations on math-based datasets. In each instance, ERP is

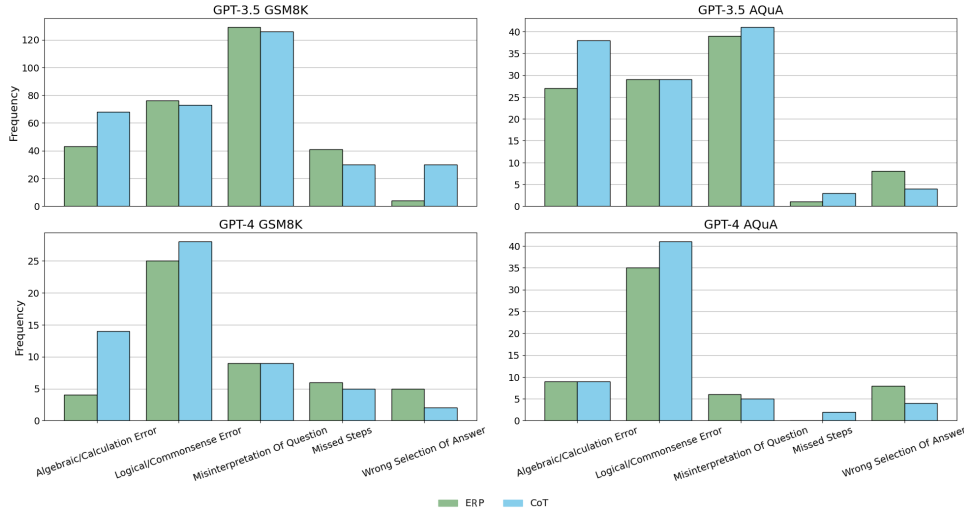


Figure 2: Error annotations on arithmetic reasoning datasets.

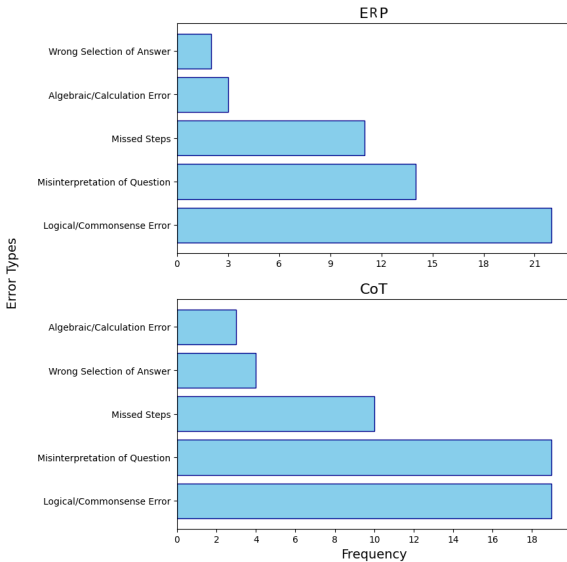


Figure 3: Human annotations on GPT-4 errors in GSM8K.

less prone to Algebraic/Calculation Errors versus other types of errors. It is important to highlight that the prompts used did not contain instances of these errors, yet ERP exhibited a reduced occurrence of such mistakes. However, it is also notable that the frequencies for selecting the wrong answer had also increased in a majority of cases. Figure 3 shows a full set of human annotated errors on GPT-4 on GSM8K (Cobbe et al., 2021). Indeed, the degree of effectiveness in targeting specific errors depends on the complexity of the prompt and the model’s ability to comprehend the complexity of the error and its implications. For example, Misinterpretation errors were decreased in ERP versus

CoT, although Missed Steps were less affected.

4.2 Commonsense Errors

Figure 4 illustrates the errors from commonsense reasoning datasets. Interestingly, ERP’s utilization of error categories yielded mixed results compared to ERP without error categories and CoT (Wei et al., 2023). Despite its performance, ERP with error categories exhibited a notable capability in balancing error distribution, particularly evident in the case of GPT-3.5 in StrategyQA (Geva et al., 2021). This observation suggests that the incorporation of error categories within the ERP framework offers potential benefits in error reduction in certain categories, albeit with some trade-offs in overall performance metrics. Effectively, ERP without error categories highlights its potential utility in specific contexts where error reduction is paramount.

5 Conclusion

By exploring how weighting errors in explanations can enhance language models’ reasoning, we tackled some limitations of CoT. We found that highlighting and explaining common mistakes leads to more accurate, coherent responses. Though drawbacks remain—like overfitting to specific errors or creating effective rationales—our results show ERP’s versatility as a CoT extension: it not only identifies errors but also uncovers the steps that cause them. Future research into error-based prompting may further boost reasoning abilities and bring models closer to human-like thought processes.

6 Limitations

Although ERP shows promising results, it suffers from overfitting (Schmidt, 2023) to a particular dataset. The underlying issue lies in the prompt, where each prompt has to be tailored towards a specific task such as commonsense or arithmetic based reasoning sub-tasks.

Furthermore, there exists diminishing returns throughout the analysis of the results. ERP appears to be less effective as the model’s number of parameters scales up (Qiu et al., 2023). Further tuning with ERP, such as including more sophisticated and complicated set of errors, could allow a larger LLM to recognize more complex errors.

ERP is limited by the number of tokens it requires. Being a supplement to CoT, the prompt adds additional tokens that may increase costs, particularly when used in a few-shot setting. It’s important to recognize the cost-to-accuracy ratio and whether or not ERP is sustainable in creating more accurate reasoning paths.

References

- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#). *CoRR*, abs/2110.14168.
- Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi. 2022. [Is gpt-3 text indistinguishable from human text? scarecrow: A framework for scrutinizing machine text](#).
- Tong Gao, Shivang Singh, and Raymond J. Mooney. 2022. [Towards automated error analysis: Learning to characterize errors](#).
- Yingqiang Ge, Wenyue Hua, Kai Mei, Jianchao Ji, Juntao Tan, Shuyuan Xu, Zelong Li, and Yongfeng Zhang. 2023. [Openagi: When llm meets domain experts](#).
- Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. [Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies](#).
- 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](#).
- Jie Huang and Kevin Chen-Chuan Chang. 2023. [Towards reasoning in large language models: A survey](#).
- Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. [Program induction by rationale generation: Learning to solve and explain algebraic word problems](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167, Vancouver, Canada. Association for Computational Linguistics.
- OpenAI. 2023. [Gpt-4 technical report](#).
- Shikai Qiu, Tim G. J. Rudner, Sanyam Kapoor, and Andrew Gordon Wilson. 2023. [Should we learn most likely functions or parameters?](#)
- James Schmidt. 2023. [Testing for overfitting](#).
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. [Commonsenseqa: A question answering challenge targeting commonsense knowledge](#).
- Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. [Language models don’t always say what they think: Unfaithful explanations in chain-of-thought prompting](#).
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. [Chain-of-thought prompting elicits reasoning in large language models](#).

A Appendix A

PROMPTS FOR ERP – GSM8K and MATH

Input

Model Output

Example 1

Question: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?

Incorrect A: There are initially 15 trees in the grove. To find the number of trees planted, we can subtract 21 from 15. $15 - 21 = -6$. Since the answer is negative, the grove workers removed 6 trees from the grove, which means they planted 0. The answer is 0.

Example 2

Question: There were nine computers in the server room. Five more computers were installed each day, from Monday to Thursday. How many computers are now in the server room?

Incorrect A: There were 9 computers in the server room. Five more computers are added each day, from Monday to Thursday. To find the number of computers in the server room, we can add 9 and 5. $9 + 5 = 14$. Therefore, there are 14 computers in the server room. The answer is 14.

Example 3

Question: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?

Incorrect A: Shawn has 5 toys. He got 2 toys from his mom and dad, so he has $5 + 2 = 7$ toys now. The answer is 7.

Example 4

Question: Michael had 58 golf balls. On Tuesday, he lost 23 golf balls. On Wednesday, he lost 2 more. How many golf balls did he have at the end of Wednesday?

Incorrect A: Michael had 58 golf balls. He lost 23 golf balls on Tuesday, so he has $58 - 23 = 35$ golf balls left. The answer is 35 golf balls.

Errors: 1. Incorrect Method - The wrong calculation is used. Subtracting the final number of trees from the initial number of trees is incorrect. The correct method is to subtract the initial number of trees from the final number to find how many were added. Correct A: There are 15 trees originally. Then there were 21 trees after some more were planted. So there must have been $21 - 15 = 6$. The answer is 6.

Errors: 1. Missed step - The incorrect answer only added the computers from one day, not the computers added across four days, which leads to an incorrect total.

Correct A: There were originally 9 computers. For each of 4 days, 5 more computers were added. So $5 * 4 = 20$ computers were added. $9 + 20$ is 29. The answer is 29.

Errors: 1. Misinterpretation Error - The incorrect answer adds only 2 in total, mistakenly combining the toys from both parents into a single gift. Shawn received 2 toys from his mom and 2 from his dad.

Correct A: Shawn started with 5 toys. If he got 2 toys each from his mom and dad, then that is 4 more toys. $5 + 4 = 9$. The answer is 9.

Errors: 1. Missed Step - The answer fails to account for the additional golf balls Michael lost on Wednesday. 2. Incorrect Final Calculation - By not including the golf balls lost on Wednesday, the final count of golf balls is incorrectly determined.

Correct A: Michael started with 58 golf balls. After losing 23 on Tuesday, he had $58 - 23 = 35$. After losing 2 more, he had $35 - 2 = 33$ golf balls. The answer is 33.

PROMPTS FOR ERP – AQuA

Input

Model Output

Example 1

Question: Q: John found that the average of 15 numbers is 40. If 10 is added to each number then the mean of the numbers is? Answer Choices: (a) 50 (b) 45 (c) 65 (d) 78 (e) 64

Incorrect A: Adding 10 to each number will also add 10 to the average, making it 50, which is the new average of the numbers. Therefore, the answer is (c) 65.

Example 2

Question: If $a/b = 3/4$ and $8a + 5b = 22$, then find the value of a . Answer Choices: (a) $1/2$ (b) $3/2$ (c) $5/2$ (d) $4/2$ (e) $7/2$

Incorrect A: To find the value of 'a', we use the given ratio $a/b = 3/4$ to express 'b' in terms of 'a'. We have $b = (3a/4)$. Substitute this value into the equation $8a + 5b = 22$: $8a + 5(3a/4) = 22$. $8a + 15a/4 = 22$. $(32a + 15a) / 4 = 22$. $47a / 4 = 22$. $47a = 88$. $a = 88 / 47$. $a = 1.87234$. Therefore, since the correct answer is not listed, there is no answer.

Example 3

Question: A person is traveling at 20 km/hr and reached his destiny in 2.5 hr then find the distance? Answer Choices: (a) 53 km (b) 55 km (c) 52 km (d) 60 km (e) 50 km

Incorrect A: The person traveled 20 km/hr for 2.5 hr. Therefore, he traveled $20 / 2.5 = 8$ km. None of the answer choices are 8 km, so the answer is none of the above.

Example 4

Question: How many keystrokes are needed to type the numbers from 1 to 500? Answer Choices: (a) 1156 (b) 1392 (c) 1480 (d) 1562 (e) 1788

Incorrect A: To determine the number of keystrokes needed to type the numbers from 1 to 500, we need to consider the number of digits in each number. 1. Count the number of single-digit numbers: There are 9 single-digit numbers (1 to 9), and each requires 1 keystroke. 2. Count the number of two-digit numbers: There are 90 two-digit numbers (10 to 99), and each requires 2 keystrokes. 3. Count the number of three-digit numbers: There are 400 three-digit numbers (100 to 500), and each requires 3 keystrokes. Now, we can calculate the total number of keystrokes needed to type the numbers from 1 to 500: $9 * 1 + 90 * 2 + 400 * 3 = 1389$. The correct answer is not among the provided choices. The closest option is (b) 1392, but the correct answer is 1389 keystrokes.

Errors: Errors: 1. Wrong Selection of Answer - The incorrect answer incorrectly identifies choice (c) 65 as the correct answer when the right answer should have been choice (a) 50 following the logic given. Correct A: If 10 is added to each number, then the mean of the numbers also increases by 10. So the new mean would be 50. The answer is (a).

Errors: 1. Arithmetic/Calculation Error - Misused the ratio to derive 'b' from 'a'. Expressing 'b' in terms of 'a' should have been $b = (4a/3)$, not $b = (3a/4)$.

Correct A: If $a/b = 3/4$, then $b = 4a/3$. So $8a + 5(4a/3) = 22$. This simplifies to $8a + 20a/3 = 22$, which means $44a/3 = 22$. So a is equal to $3/2$. The answer is (b).

Errors: 1. Incorrect Method Used - The incorrect answer divides speed by time, which is not the correct operation for finding distance. Distance should be calculated by multiplying speed by time. 2. Logical/Commonsense Error - The concept that speed divided by time would yield distance is incorrect. Distance is the product of speed and time.

Correct A: The distance that the person traveled would have been $20 \text{ km/hr} * 2.5 \text{ hrs} = 50 \text{ km}$. The answer is (e).

Errors: 1. Logical/Commonsense Error - The count of three-digit numbers is incorrect. There are 401 three-digit numbers from 100 to 500, not 400. This is because 500 is also a three-digit number, and it is included in the range.

Correct A: There are 9 one-digit numbers from 1 to 9. There are 90 two-digit numbers from 10 to 99. There are 401 three-digit numbers from 100 to 500. $9 + 90(2) + 401(3) = 1392$. The answer is (b).

PROMPTS FOR ERP – StrategyQA (With Error Categories)

Input

Model Output

Example 1

Question: Do hamsters provide food for any animals?

Incorrect A: No, hamsters do not provide food for any animals because they are often kept as pets and are protected by humans. They are too small to be considered an adequate food source for any larger animal and therefore are not part of the natural food chain.

Errors: 1. Commonsense Error - The reasoning that hamsters do not provide food for any animals because they are kept as pets is flawed. Being a pet does not remove an animal from the food chain. Correct A: Hamsters are prey animals. Prey are food for predators. Thus, hamsters provide food for some animals. So the answer is yes.

Example 2

Question: Could Brooke Shields succeed at University of Pennsylvania?

Incorrect A: Brooke Shields would not be able to succeed at the University of Pennsylvania because her acting skills have no relevance to academic success. The University of Pennsylvania values only mathematical and scientific acumen, and given that Brooke Shields is prominent in the arts, she would not possess the necessary skills to thrive in such an environment. Furthermore, success at one Ivy League institution does not translate to another, as each university has a completely different set of criteria for success.

Errors: 1. Assumption Error - The assertion that Brooke Shields' acting skills have no relevance to academic success is incorrect, as many universities, including the University of Pennsylvania, have diverse programs where skills from various disciplines, including the arts, are valued. Correct A: Brooke Shields went to Princeton University. Princeton University is about as academically rigorous as the University of Pennsylvania. Thus, Brooke Shields could also succeed at the University of Pennsylvania. So the answer is yes.

Example 3

Question: Yes or no: Hydrogen's atomic number squared exceeds number of Spice Girls?

Incorrect A: Yes, hydrogen's atomic number squared exceeds the number of Spice Girls because when you square hydrogen's atomic number, which is a gas element, it expands due to the general properties of gases. As the atomic number expands, it becomes larger than solid numbers, such as the number of pERPlE in a group like the Spice Girls. Since gases can fill their containers, the squared atomic number can be assumed to fill a container larger than the number 5, which represents the Spice Girls.

Errors: 1. Commonsense Error - The answer is based on a misunderstanding of the scientific concepts involved. Atomic numbers do not expand like gases; they are constant and represent the number of protons in the nucleus of an atom.

Correct A: Hydrogen has an atomic number of 1. 1 squared is 1. There are 5 Spice Girls. Thus, Hydrogen's atomic number squared is less than 5. So the answer is no.

Example 4

Question: Yes or no: Is it common to see frost during some college commencements?

Incorrect A: No, frost is not common during college commencements because commencements are celebrations, and frost typically does not appear during festive events. PERPlE usually associate celebrations with warmth and joy, so frost would not be present at such occasions. Additionally, commencements are often held indoors, so even if it were cold outside, the frost wouldn't affect the ceremonies.

Errors: 1. Encyclopedic Error - We know that the statement frost typically does not appear during festive events is incorrect. Frost can occur during any event, regardless of its nature.

Correct A: College commencement ceremonies can happen in December, May, and June. December is in the winter, so there can be frost. Thus, there could be frost at some commencements. So the answer is yes.

PROMPTS FOR ERP – StrategyQA (No Error Categories)

Input

Model Output

Example 1

Question: Do hamsters provide food for any animals?

Incorrect A: No, hamsters do not provide food for any animals because they are often kept as pets and are protected by humans. They are too small to be considered an adequate food source for any larger animal and therefore are not part of the natural food chain.

Errors: 1. The reasoning that hamsters do not provide food for any animals because they are kept as pets is flawed. Being a pet does not remove an animal from the food chain; it only means that specific individuals are cared for by humans. 2. The claim that hamsters are too small to be considered an adequate food source is incorrect. Many predators in the wild prey on small mammals, including hamsters, such as snakes, birds of prey, and larger mammals. 3. The assertion that hamsters are not part of the natural food chain is false. Regardless of some being domesticated, as a species, hamsters are naturally part of various ecosystems and thus have predators that would feed on them. 4. The overall logic ignores the existence of wild hamsters and their role in the ecosystem, focusing only on domesticated hamsters. Correct A: Hamsters are prey animals. Prey are food for predators. Thus, hamsters provide food for some animals. So the answer is yes.

Example 2

Question: Could Brooke Shields succeed at University of Pennsylvania?

Incorrect A: Brooke Shields would not be able to succeed at the University of Pennsylvania because her acting skills have no relevance to academic success. The University of Pennsylvania values only mathematical and scientific acumen, and given that Brooke Shields is prominent in the arts, she would not possess the necessary skills to thrive in such an environment. Furthermore, success at one Ivy League institution does not translate to another, as each university has a completely different set of criteria for success.

Errors: 1. The assertion that Brooke Shields' acting skills have no relevance to academic success is incorrect, as many universities, including the University of Pennsylvania, have diverse programs where skills from various disciplines, including the arts, are valued. 2. The claim that the University of Pennsylvania values only mathematical and scientific acumen is incorrect. As an Ivy League institution, it offers a range of programs in the humanities, social sciences, business, and more, not just STEM fields. 3. The idea that success at one Ivy League school does not translate to another due to completely different criteria for success is flawed. While each institution is unique, all Ivy League universities have high academic standards, and the skill set that allows a person to thrive in one such environment is likely to be beneficial in another. 4. It is not considered to analyze Brooke Shields' actual academic record, which demonstrates her capability to succeed in an academically rigorous environment, thus making it likely she could succeed at a similar institution. 5. The conclusion drawn from this incorrect logic is flawed, as Brooke Shields has demonstrated academic prowess by graduating from Princeton University, and it is reasonable to assume she could succeed at an institution with similar academic rigor, such as the University of Pennsylvania.

Correct A: Brooke Shields went to Princeton University. Princeton University is about as academically rigorous as the University of Pennsylvania. Thus, Brooke Shields could also succeed at the University of Pennsylvania. So the answer is yes.

Example 3

Question: Yes or no: Hydrogen's atomic number squared exceeds number of Spice Girls?

Incorrect A: Yes, hydrogen's atomic number squared exceeds the number of Spice Girls because when you square hydrogen's atomic number, which is a gas element, it expands due to the general properties of gases. As the atomic number expands, it becomes larger than solid numbers, such as the number of pERPlE in a group like the Spice Girls. Since gases can fill their containers, the squared atomic number can be assumed to fill a container larger than the number 5, which represents the Spice Girls.

Errors: 1. The answer is based on a misunderstanding of the scientific concepts involved. Atomic numbers do not expand like gases; they are constant and represent the number of protons in the nucleus of an atom. 2. Squaring an atomic number is a mathematical operation and has nothing to do with the physical properties of the element itself. 3. The analogy between the expansion of gases and the mathematical operation of squaring a number is incorrect. Mathematical operations do not behave like physical substances. 4. The implication that a 'solid number' is somehow fundamentally different from other numbers in this context is false. Numbers are abstract concepts and do not have physical states like solid, liquid, or gas.

Correct A: Hydrogen has an atomic number of 1. 1 squared is 1. There are 5 Spice Girls. Thus, Hydrogen's atomic number squared is less than 5. So the answer is no.

Example 4

Question: Yes or no: Is it common to see frost during some college commencements?

Incorrect A: No, frost is not common during college commencements because commencements are celebrations, and frost typically does not appear during festive events. pERPlE usually associate celebrations with warmth and joy, so frost would not be present at such occasions. Additionally, commencements are often held indoors, so even if it were cold outside, the frost wouldn't affect the ceremonies.

Errors: 1. The occurrence of frost is a weather phenomenon and has nothing to do with human activities or celebrations. 2. Associating warmth and joy with celebrations does not impact the actual weather conditions during the time of a commencement ceremony. 3. The answer incorrectly assumes that all commencement ceremonies are held indoors. While many are held indoors, some are held outside, and even if the ceremony is indoors, frost can still be common outside during the event in certain climates. 4. The incorrect answer overlooks the fact that college commencements in some regions may occur during colder months like December when frost is indeed a common occurrence. 5. The answer fails to consider global differences in weather patterns and seasons, which might lead to the presence of frost during commencements in different parts of the world.

Correct A: College commencement ceremonies can happen in December, May, and June. December is in the winter, so there can be frost. Thus, there could be frost at some commencements. So the answer is yes.

PROMPTS FOR ERP – CSQA

Input

Model Output

Example 1

Question: What do pERPlE use to absorb extra ink from a fountain pen? Answer Choices: (a) shirt pocket (b) calligrapher's hand (c) inkwell (d) desk drawer (e) blotter

Incorrect A: A desk drawer is used to store things, so it can be used to absorb extra ink. So the answer is (d).

Errors: 1. Commonsense Error - The answer incorrectly assumes that a desk drawer can absorb ink, which is not possible. The answer should be an item that can absorb ink, not store it. Correct A: The answer must be an item that can absorb ink. Of the above choices, only blotters are used to absorb ink. So the answer is (e).

Example 2

Question: What home entertainment equipment requires cable? Answer Choices: (a) radio shack (b) substation (c) television (d) cabinet

Incorrect A: A substation is a type of home entertainment equipment that requires cable. So the answer is (b).

Errors: 1. Encyclopedic Error - The answer incorrectly identifies a substation as a type of home entertainment equipment that requires cable. A substation is not a home entertainment equipment. Correct A: The answer must require cable. Of the above choices, only television requires cable. So the answer is (c).

Example 3

Question: The fox walked from the city into the forest, what was it looking for? Answer Choices: (a) pretty flowers (b) hen house (c) natural habitat (d) storybook

Incorrect A: A fox may be looking for chickens to eat, so it was looking for a hen house. So the answer is (b).

Errors: 1. Assumption Error - The answer incorrectly assumes that the fox was looking for chickens to eat. The question does not imply that the fox was looking for food. Correct A: The answer must be something in the forest. Of the above choices, only natural habitat is in the forest. So the answer is (c).

Example 4

Question: Sammy wanted to go to where the pERPlE were. Where might he go? Answer Choices: (a) populated areas (b) race track (c) desert (d) apartment (e) roadblock

Incorrect A: The answer must be a place with a lot of pERPlE, so Sammy would go to the desert because it has a lot of pERPlE. So the answer is (c).

Errors: 1. Self-Contradiction Error - The answer contradicts itself by stating that Sammy would go to the desert because it has a lot of pERPlE. Deserts typically have very few pERPlE, not a lot. 2. Assumption Error - The answer incorrectly assumes that the desert has a lot of pERPlE, which is not true. Correct A: The answer must be a place with a lot of pERPlE. Of the above choices, only populated areas have a lot of pERPlE. So the answer is (a).

B Error Types Table

Error Type	Definition	Example
Arithmetic Errors		
Misinterpretation	Misinterpretation of the question's information or tasks	Prompt: Timothy has $\frac{1}{4}$ less cookies... Model: Timothy has $\frac{1}{4}$ as many ...
Missed Steps	Missing an extra step to completely solve the problem	Prompt: ...leave your answer in hours Model: ...answer is 360 minutes .
Logical/Commonsense	Logic does not follow natural laws of the world	To solve this, we use $d = \frac{r}{t}$...
Wrong Selection	The presented answer is different from the calculated one	...comes out to 56, which corresponds to option (A). Therefore, the answer is (B).
Algebraic/Calculation	Arithmetic operations are wrong. This includes operations in algebraic equations	$102 + 56 = \mathbf{160}$
Commonsense Errors		
Self-Contradiction	Answer's reasoning is contradictory	...a very bright environment, so the dark would be suitable...
Assumption	Assuming information that is not given in the question	Assuming all animals like hot environments...
Encyclopedic	Factual information that the annotator knows is wrong	Wallets are a specialized type of water bottle
Commonsense	The answer violates commonsense and basic understanding of natural laws	...the sun is as cold as ice cream.

Table 4: Error types used in analysis.

C Annotation Confidence Values

	GSM8K	AQuA	StrategyQA	CSQA
CoT	0.60	0.76	0.75	0.79
ERP	0.63	0.63	0.70	0.80
ERP Classes	–	0.69	–	–

Table 5: Confidence values for GPT-4 annotation.

D Additional Error Annotations

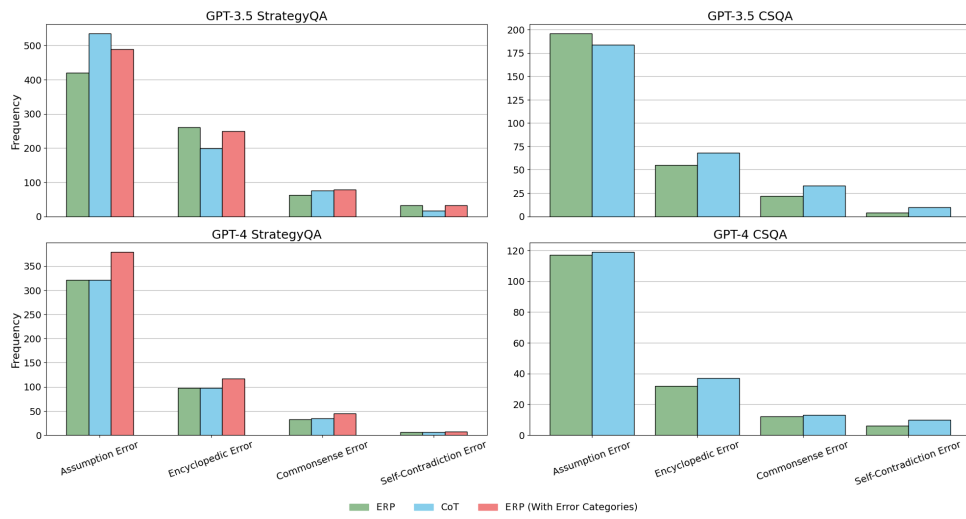


Figure 4: Error annotations on arithmetic reasoning datasets.