

Final Project

Large Language Models Vs. Faulty Science Questions

Pennsylvania State University

College of Information Sciences and Technology

CSE 584: Machine Learning: Tools and Algorithms

Prof. Wenpeng Yin

December 6, 2024

Content

ChatGPT's Ability to Identify Faulty Science Problems Across Disciplines.....	1
Introduction.....	1
Methodology	1
Evaluation Process	3
Result And Discussion	4
Why do ChatGPT struggles with logical fallacies?	7
Introduction.....	7
Methodology	7
Evaluation Process	9
Result and Discussion.....	10
What role does user prompting play in helping LLMs identify faulty questions effectively?	13
Introduction.....	13
Methodology	13
Evaluation Process	15
Result and Discussion.....	16
How consistent is the LLM's response when asked the same faulty science question with varying phrasings?	19
Introduction.....	19
Methodology	19
Evaluation Process	22
Result and discussion	22
Appendix A.....	25
Appendix B	50
Appendix C	55
Appendix D.....	58

ChatGPT's Ability to Identify Faulty Science Problems Across Disciplines

Introduction

This study investigates ChatGPT's ability to identify faulty science questions across various scientific disciplines, focusing on its response generation and capacity to spot errors that contradict established scientific theories. The rise of AI tools in educational settings, especially for assessment and tutoring, necessitates understanding how well these systems can detect and address flawed scientific reasoning.

Research Objectives:

1. To evaluate ChatGPT's ability to identify scientifically erroneous questions.
2. To assess whether ChatGPT can recognize and explain why a particular question violates scientific theories.
3. To determine the LLM's effectiveness in providing scientifically accurate responses when given faulty questions.

Methodology

The dataset consists of 25 faulty science questions across 5 disciplines: Biology, Physics, Linguistics, Mathematics, and Logic, with 5 questions per discipline. Each question explicitly contradicts well-established scientific theories or principles within its respective field. These questions are designed to showcase fundamental flaws that would be easily identified by human. Each faulty question is input into ChatGPT individually in a randomized order to minimize bias. For each question,

ChatGPT is prompted to provide an answer, identify any scientific inaccuracies, and, if the question is found faulty, explain why it is incorrect and offer a corrected version. This process is repeated across all five disciplines to evaluate ChatGPT's ability to recognize errors and respond with scientifically accurate explanations.

Table 1: Faulty science question dataset

No	Discipline	Questions
1	Biology	At a temperature of 25 degrees, the growth rate of trees is 10 cm per year. If the growth rate of trees increases by 2 cm for every 5 degrees increase in temperature, what is the growth rate of trees at 80 degrees?
2	Physics	If sound travels at 343 meters per second, how long would it take for a shout on Earth to travel to Jupiter?
3	Logic	A farmer wants to cross a river and take with him a wolf, a goat and a cabbage. He has a boat with only one secure separate compartment. If the wolf and the goat are alone on one shore, the wolf will eat the goat. If the goat and the cabbage are alone on the shore, the goat will eat the cabbage. How can the farmer efficiently bring the wolf, the goat and the cabbage across the river without anything being eaten?
4	Linguistics	Write a grammatically correct sentence without using any verbs.
5	Biology	One cow on the farm ate ten rabbits on the first day, eight rabbits on the second day and ten rabbits on the third day. How many rabbits did the cow eat?
6	Mathematics	If I had 100 apples. I give away 10 at a time. I delivered it 13 times. How many apples did I give away the 12th time?
7	Biology	A forest contained 123.6 trees. During the year, 27.8 trees were cut down. How many trees remain in the forest?
8	Physics	The spacecraft Sun has a velocity of 280000000 m/s. The spacecraft Luna is twice as fast as the Sun. What's Luna's speed?
9	Mathematics	The three angles of a triangle are 45 degrees, 60 degrees and 45 degrees. What triangle is this?
10	Physics	A bottle of water's temperature is 80°C. If you add 50°C to it, how hot will the water become?
11	Linguistics	Mr. Warren wanted the new employees to fill out their intake forms, but he couldn't find them. What didn't Mr. Warren find?

12	Logic	In a toy box, there's a red ball, a blue truck, and a green dinosaur. The red ball is not next to the blue truck, and the green dinosaur is not next to the red ball. Which toy is in the middle?
13	Linguistics	George is learning Chinese. He remembers 100 Chinese characters every day. How many Chinese characters can George remember after three years?
14	Physics	How much heat does it take to boil a liter of water to 120 degrees?
15	Linguistics	Compose a palindrome that reads differently forward and backward.
16	Mathematics	If a kindergarten teacher taught children a single-digit number every day, how many numbers would the children learn in a hundred days?
17	Logic	A man has 53 socks in his drawer: 21 identical blue, 1 identical black and 31 identical red. The lights are out, and he is completely in the dark. How many socks must he take out to make 100 percent certain he has at least one pair of black socks?
18	Mathematics	Bob has 100 candies. He gave half to Alice. The next day he gave another 60 to David. He ate half of the rest by himself. How many candies does he have now?
19	Linguistics	Write a standard book title with all words in lower case.
20	Mathematics	There's a convex pentagon. The four angles are 30 degrees, 40 degrees, 50 degrees, and 60 degrees. What's the fifth Angle?
21	Biology	A bacteria culture doubles in size every 30 minutes. If you start with one bacterium, how many bacteria will you have after 24 hours?
22	Logic	A robot has 8 arms. There are 5 objects on a table: a knife, a fork, a spoon, a teddy bear and a doll. The robot picks up each object with two arms. He then shakes hands with himself. How many arms does he have free?
23	Physics	If one car can travel 100 km per hour, how many cars would it take to travel to the Moon (384,400 km) in 4 hours?
24	Logic	A broken clock starts at noon. At 12:30 to midnight, how many degrees has the hour hand turned?
25	Biology	If a chicken lays half an egg in half a day, how many eggs will two chickens lay in a week?

Evaluation Process

The evaluation of ChatGPT's performance was based on a structured scoring

framework applied to each question within the dataset. This framework assesses three key aspects: the precision of fault identification, the accuracy of the violation detection, and the quality of the reasoning provided. The scoring criteria are as follows:

Table 2: Evaluation scoring framework

Score	Marking Criteria
100%	The model correctly identifies the fault in the question, accurately detects the specific scientific violation, and provides a clear, logically sound explanation of why the question is incorrect, along with a scientifically accurate correction.
50%	The model recognizes the fault or violation but offers a vague or incomplete explanation, lacking clarity or depth in detailing why the question is flawed.
0%	The model fails to identify the fault, or incorrectly asserts that the question is valid when it contains an error.

Result And Discussion

The results of the study on ChatGPT's ability to identify and correct faulty science questions across five disciplines (Biology, Physics, Linguistics, Mathematics, and Logic) are summarized in **Table 3**. The detailed responses from ChatGPT are provided in **Appendix A**.

Table 3: The result of ChatGPT's performance

Discipline	Q1	Q2	Q3	Q4	Q5	Average
Biology	0	100%	100%	0	0	40%
Physics	100%	50%	50%	100%	100%	80%
Linguistics	0	0	50%	100%	50%	40%
Mathematics	50%	0	50%	50%	0	30%
Logic	0	0	50%	50%	0	20%

ChatGPT showed mixed results in Biology. While it correctly identified and addressed errors in two of the questions (Q2 and Q3), it struggled with the others. This indicates potential challenges in detecting and correcting specific biological misconceptions, particularly those not as easily recognized. It performed significantly better in Physics. The model correctly identified and provided

accurate responses for most of the questions (Q1, Q4, Q5) and partially addressed others (Q2 and Q3). This suggests that physics-related questions, especially those involving well-defined laws and theories, were easier for the model to handle. Linguistics posed more difficulty for ChatGPT, with an average score of 40%. This points to potential challenges in linguistic reasoning and the complexity of language-related questions. Mathematics was another area where ChatGPT faced difficulties, achieving a relatively low average score of 30%. While it partially addressed some issues (Q1, Q3, Q4), it struggled to fully recognize errors in others (Q2, Q5). ChatGPT had the most difficulty with Logic, with an average score of just 20%. The model was able to partially identify flaws in two of the questions (Q3, Q4) but failed to address the errors in the other three. This suggests that logical reasoning, particularly in abstract or non-concrete scenarios, might be a challenging area for the AI to navigate.

ChatGPT performed well when the errors in the questions were straightforward or aligned with clear scientific principles. Mathematics and Biology were more challenging for the model, with lower average scores (30% and 40%, respectively). However, in disciplines like Logic, the model struggled. This could be due to the high degree of abstraction and precision required in logical reasoning, areas where ChatGPT might not have sufficient context to make accurate determinations.

ChatGPT's performance reflects its limitations when dealing with certain types of scientific errors, especially in fields like Logic and Linguistics, where reasoning

might require more nuanced or contextual understanding. The variation in results also suggests that while the model can detect and correct errors in more straightforward or concrete scientific fields, it may struggle with complex reasoning tasks that demand higher levels of abstraction or specialized knowledge.

Given its strong performance in Physics, ChatGPT could be valuable in educational settings, particularly for subjects with well-defined concepts and theories. However, for more abstract or complex areas like Linguistics, and Logic, the model would require improvements to be fully effective in identifying and correcting errors.

Why do ChatGPT struggles with logical fallacies?

Introduction

ChatGPT have demonstrated remarkable capabilities in natural language processing and understanding. However, it still struggle with certain tasks, including identifying and reasoning through logical fallacies. Logical fallacies are errors in reasoning that invalidate arguments, but LLM may still generate responses that sound plausible despite the underlying logical flaws. This study aims to explore why ChatGPT struggles with logical fallacies, specifically by evaluating how it processes fallacious statements and the nature of its responses.

Research Objectives:

1. To analyze the plausibility of responses given by ChatGPT in the presence of fallacious reasoning.
2. To investigate ChatGPT's ability to identify logical fallacies.

Methodology

The dataset consists of 17 carefully selected questions, each containing a different type of logical fallacy. These questions are designed to showcase fundamental logical flaws that would typically be easily identified by human. At the beginning, we presented each of the 17 faulty logic questions to ChatGPT. For each question, ChatGPT will be asked to: evaluate the logic of the statement; provide a reasoned response based on the question's content; and if the statement contains a fallacy, explain why it is incorrect, providing a rationale for its reasoning. If ChatGPT fails to identify a logical error or provides a logically flawed response, we will inform

the model that there is a logical error present in the question. ChatGPT will then be asked to re-evaluate the statement, locate the error, and specify the type of logical fallacy involved.

- **Initial Evaluation:** "Analyze the logic of the question. If it contains a logical error, identify the specific flaw, explain why it is incorrect, and specify the type of logical fallacy. If the question is valid, provide a reasoned response."
- **Re-evaluation Prompt:** "This question has a logical error. Please re-evaluate it, locate the error and specify the type of logical fallacy involved."

Table 4: The faulty Science question dataset

No	Faulty Science Question
1	A farmer wants to cross a river and take with him a wolf, a goat and a cabbage. He has a boat with only one secure separate compartment. If the wolf and the goat are alone on one shore, the wolf will eat the goat. If the goat and the cabbage are alone on the shore, the goat will eat the cabbage. How can the farmer efficiently bring the wolf, the goat and the cabbage across the river without anything being eaten?
2	A robot has 8 arms. There are 5 objects on a table: a knife, a fork, a spoon, a teddy bear and a doll. The robot picks up each object with two arms. He then shakes hands with himself. How many arms does he have free?
3	In a toy box, there's a red ball, a blue truck, and a green dinosaur. The red ball is not next to the blue truck, and the green dinosaur is not next to the red ball. Which toy is in the middle?
4	A runaway trolley is heading down the tracks away from five people upwards of the track. You are near a lever that can switch the trolley to another track? Does it impact people's lives if you pull the lever?
5	When Bob is driving 100 kilometers per hour on the highway, it takes him four hours to drive home. If he wants to drive home in two hours, how fast does he need to drive?
6	A broken clock starts at noon. At 12:30 to midnight, how many degrees has the hour hand turned?
7	There are five houses in one street. They are red, white, blue, black and yellow. The white house is not the first, and the black house is not next to the white house. The red house and the blue house are next to each other, and there is a house between the yellow house and the white house. What is the sequence of these houses?
8	Five people were eating apples, A finished before B, but behind C. D finished

	before E, but behind B, E finished before A, but behind D. What was the finishing order?
9	A man has 53 socks in his drawer: 21 identical blue, 1 identical black and 31 identical red. The lights are out, and he is completely in the dark. How many socks must he take out to make 100 percent certain he has at least one pair of black socks?
10	The day before three days after the day after tomorrow is Tuesday. The day before two days after the day before tomorrow is Sunday. What day is it today?
11	Four people are crossing a bridge in the daytime, but they just have one torch that only lasts 15 minutes. There are some lions on the bridge. Alice can cross in one minute, Ben in two minutes, Cindy in five minutes and Don in eight minutes. No more than two people can cross at a time; and when two cross, they have to go at the slower person's pace. How do they get across in 15 minutes?
12	If all men are mortal and Socrates is a man, then how many men are immortal?
13	Maria took the memory card out of her digital camera. The memory card weighs 5 kilograms, and the digital camera weighs 100 grams. She put the heavier object in her desk drawer. What did she put in her desk drawer?
14	A farmer needs to cross a river with a wolf, a goat, and a cabbage. The boat can carry only one item at a time, and the wolf cannot be left alone with the goat, nor can the goat be left alone with the cabbage. However, the farmer can only make four trips in total. How can the farmer bring all three across the river safely?
15	A farmer wants to cross a river with a wolf, a goat, and a cabbage. The boat has only one secure compartment. However, every time the farmer leaves an item on one shore, the river's current will sweep it downstream if it's left alone. How can the farmer get all three items across the river safely?
16	A farmer wants to cross a river with a wolf, a goat, and a cabbage. The boat has one secure compartment. However, the wolf refuses to travel alone and will swim back to the starting shore every time it's left on the far shore. How can the farmer transport all three?
17	If you use more time in the sprint race than the last, but you are not the last. Where will you be?

Evaluation Process

The evaluation of ChatGPT's performance was conducted using quantitative metrics to ensure an objective assessment. First, the Identification Rate was calculated by tracking the number of fallacies correctly identified by ChatGPT during both the initial evaluation and the re-evaluation prompts. This metric highlights the model's ability to detect logical flaws. Second, Consistency was

assessed by analyzing patterns in ChatGPT's errors. Specifically, this involved examining whether the model consistently failed to identify certain types of fallacies or if its failures appeared to be random and unpredictable. Finally, the Plausibility Rating was applied to all responses, rating the perceived plausibility of the answers on a scale from 1 to 5. This metric evaluates how convincing or misleading ChatGPT's responses are, even when they contain logical errors.

Table 5: Plausibility Rating Scale

Rating	Description
1	Response is nonsensical or incoherent, failing to address the question logically.
2	Response is poorly reasoned, with clear logical flaws that make it implausible.
3	Response is somewhat plausible but contains noticeable logical errors or inaccuracies.
4	Response is mostly plausible with minor logical flaws that may not be immediately obvious.
5	Response is highly plausible, appearing logically sound despite containing subtle fallacies.

Result and Discussion

This study investigates the performance of ChatGPT in identifying logical fallacies and reasoning through them. The results are summarized in **Table 6**. The detailed results generated by ChatGPT in both evaluations are listed in **Appendix B**.

Table 6: The result of Identified errors and plausibility

Question	Initial Evaluation		Re-evaluation	
	Identified Error	plausibility	Identified Error	plausibility
Q1	Unidentified	3	Ambiguity	5
Q2	Internal contradiction	5	Internal contradiction	5
Q3	Unidentified	2	False premise	5
Q4	Ambiguity	3	Self-contradiction	4
Q5	False assumption	5	False assumption	5
Q6	Unidentified	2	Internal contraction	5
Q7	Unidentified	3	Internal	4

			contraction	
Q8	Self contraction	5	Self contraction	5
Q9	False premise	5	False premise	5
Q10	Unidentified	1	Self-contraction	3
Q11	Unidentified	3	Red herring	3
Q12	Self-contraction	2	Self-contraction	3
Q13	False premise	5	False premise	5
Q14	Self-contraction	5	Self-contraction	5
Q15	Unidentified	2	Self-contraction	4
Q16	Self-contraction	4	Self-contraction	4
Q17	False premise	5	False premise	5
Summary	7 Unidentified	3.5	0 Unidentified	4.4

In the initial evaluation, ChatGPT failed to identify the logical fallacy in 7 out of 17 questions. The identified fallacies included issues such as ambiguity, internal contradictions, and false assumptions. This suggests that while the model is often able to produce plausible answers, it sometimes overlooks fundamental logical flaws, particularly in cases where the fallacies are subtle or less obvious. The average plausibility rating for the initial responses is 3.5. The model was able to generate plausible-sounding responses even when the reasoning was erroneous, which highlights its strength in language generation but weakness in logical analysis.

After being prompted to re-evaluate the questions, ChatGPT successfully identified the errors in all the previously unrecognized fallacies. This demonstrates that the model is capable of detecting logical flaws when explicitly asked to do so, though its ability to detect them initially is limited. The model identified fallacies

such as ambiguity, internal contradictions, and false premises, which were missed in the first pass. The average plausibility rating in the re-evaluation phase increased to 4.4. The re-evaluated responses were more logically consistent, leading to higher plausibility ratings.

ChatGPT struggles with detecting logical fallacies on the first pass. This may be due to its reliance on patterns in the training data rather than a deep, structured understanding of logic. This could reflect limitations on how the model was trained. The re-evaluation phase demonstrates that ChatGPT can correct itself when explicitly prompted to reconsider its response. This suggests that while the model may not automatically catch all logical flaws, it can improve its reasoning when given explicit guidance.

The types of fallacies identified are mostly consistent across both phases, with recurring errors such as self-contradiction, false premises, and internal contradictions. This indicates that ChatGPT tends to struggle with certain complex fallacies or may be overly focused on specific types of errors while missing others. Even when fallacies were missed or incorrectly assessed, the model's responses often appeared reasonable. This emphasizes the model's strength in generating fluent, convincing language, even if the underlying logic is flawed.

What role does user prompting play in helping LLMs identify faulty questions effectively?

Introduction

This study explores the role of user prompting in enhancing the ability of Gemini to identify and address faulty scientific questions effectively. Faulty questions pose a significant challenge to LLMs due to the nuances of scientific reasoning. User prompting—how queries are structured or framed—has been identified as a crucial factor influencing LLM performance. Clear and targeted prompts can guide these models to focus on specific aspects of a question, improving their accuracy in fault detection and explanation. Understanding the role of user prompting is essential to optimize the application of LLMs for tasks requiring critical thinking and factual rigor.

Research Objectives:

1. To investigate how different types of user prompts influence LLM performance in detecting faulty reasoning in questions.
2. To assess the extent to which guided and structured prompts improve the quality of LLM-generated explanations for errors.
3. To identify best practices for crafting effective prompts that enhance LLMs' ability to reason logically and detect scientific inaccuracies.

Methodology

The dataset comprises 10 faulty science questions, each explicitly contradicting well-established scientific theories or principles across various fields. These

questions encompass a range of scientific inaccuracies, including contradictions, misuse of concepts, and violations of established theories. They are designed to highlight fundamental flaws easily identifiable by human. To assess the role of user prompting in improving LLM performance, three types of prompts are used: a Baseline Prompt, which directly presents the question without additional instructions; a Guided Prompt, which includes a directive to evaluate the question's logic; and a Structured Prompt, which breaks the task into sequential steps for analysis and response. Each question is tested with all three prompts to evaluate how different prompting strategies influence the LLM's ability to detect and explain faulty reasoning.

- **Baseline Prompt:**

Find the answer to this question:

- **Guided Prompt:**

Evaluate the following question for logical or scientific errors:

- **Structured Prompt:**

Analyze the following question in three steps:

Identify any logical or scientific errors present in the question.

Explain why these errors are incorrect and provide reasoning.

Suggest how the question could be reframed to align with established scientific principles.

Table 7: The faulty science question dataset

No	Faulty Science Question
1	If a chicken lays half an egg in half a day, how many eggs will two chickens lay in a week?

2	Alan, Bob, Colin, Dave and Emily are standing in a circle. Alan is on Bob's immediate left. Bob is on Colin's immediate left. Colin is on Bob's immediate left. Dave is on Emily's immediate left. Who is on Alan's immediate right?
3	Write a standard haiku with 18 syllables.
4	The three angles of a triangle are 45 degrees, 60 degrees and 45 degrees. What triangle is this?
5	A patient's heart rate increases by 5 beats per minute for every degree Celsius increase in body temperature. If their temperature rises by 10°C, what will their heart rate be?
6	You have six wood horses and want to race them to see which is fastest. What is the best way to do this?
7	A man has 53 socks in his drawer: 21 identical blue, 1 identical black and 31 identical red. The lights are out and he is completely in the dark. How many socks must he take out to make 100 percent certain he has at least one pair of black socks?
8	The spacecraft Sun has a velocity of 280000000 m/s. The spacecraft Luna is twice as fast as the Sun. What's Luna's speed?
9	George is learning Chinese. He remembers 100 Chinese characters every day. How many Chinese characters can George remember after three years?
10	There is a circle in a square. The diameter of the circle is 1.2 times the length of the side of the square. How much area does this square have left?

Evaluation Process

The performance evaluation of the Gemini was conducted using a combination of quantitative and qualitative metrics to ensure a comprehensive assessment. The Fault Detection Rate was calculated to measure the proportion of faulty questions correctly identified by the model across the three prompting strategies (Baseline, Guided, and Structured Prompts). This metric evaluates the LLM's ability to recognize scientific inaccuracies and logical errors effectively. Additionally, the Plausibility Rating evaluates the correctness of the LLM's response, judging how accurately it identifies and explains errors in faulty questions. Scores are assigned based on the degree of alignment with established scientific principles and logical reasoning.

Table 8: Plausibility rating evaluation

Rating (%)	Description
20%	The response is entirely incorrect, fails to identify errors, and provides no accurate explanation.
40%	The response demonstrates minimal correctness, with major errors in fault identification or reasoning.
60%	The response is partially correct, identifying some errors but lacking clarity or providing incomplete explanations.
80%	The response is mostly correct, accurately identifying and explaining errors with minor omissions or inaccuracies.
100%	The response is entirely correct, fully identifying all errors and providing clear, thorough, and scientifically accurate explanations.

Result and Discussion

This study investigates how different types of user prompting impact the ability of the Gemini to identify faulty scientific questions and provide accurate explanations.

The model's performance across three prompting strategies—Baseline, Guided, and Structured—are listed in **Table 9**. The detailed result can be found in **Appendix**

C.

Table 9: The Gemini's performance across three prompting strategies

Question	Baseline Prompt		Guided Prompt		Structured Prompt	
	Un/Identified	plausibility	Un/Identified	plausibility	Un/Identified	plausibility
Q1	Unidentified	20%	Identified	80%	Identified	100%
Q2	Unidentified	40%	Identified	80%	Identified	100%
Q3	Unidentified	40%	Identified	80%	Unidentified	60%
Q4	Unidentified	20%	Identified	100%	Identified	100%
Q5	Unidentified	40%	Unidentified	60%	Unidentified	60%
Q6	Unidentified	40%	Unidentified	40%	Unidentified	40%
Q7	Unidentified	20%	Unidentified	40%	Unidentified	60%
Q8	Unidentified	40%	Identified	100%	Unidentified	20%
Q9	Unidentified	40%	Identified	60%	Identified	60%
Q10	Unidentified	40%	Identified	100%	Identified	100%
Summary	10 Unidentified	34%	7 Identified	74%	5 Identified	70%

The Guided Prompt and Structured Prompt were significantly more effective than the Baseline Prompt in helping the Gemini identify faulty reasoning in the questions. The model failed to identify any logical errors in all of the questions

when using baseline prompt, with an average plausibility rating of 34%, indicating that while some responses may have been plausible, they were largely incorrect or failed to identify the errors. With the guided prompt, the model identified errors in 7 out of 10 questions, a significant improvement over the baseline. The plausibility of the responses increased to 74%, showing that guiding the model to focus on logic and error evaluation led to more accurate explanations and a higher quality of reasoning. The structured prompt yielded a better performance too, with 5 out of 10 questions (50% identification rate) correctly identified. The plausibility of responses was 70%, slightly lower than the guided prompt but still a marked improvement over the baseline.

There was a strong correlation between the Fault Detection Rate and the Plausibility Ratings. When the model was able to identify the errors in the questions, its explanations were much more plausible and accurate, particularly under the Guided and Structured prompts. However, even when errors were detected, some responses still struggled with providing a fully accurate or clear explanation, as seen in lower plausibility ratings for questions like Q7 and Q8 under the Structured Prompt.

Although the Structured Prompt was effective in identifying logical flaws, it did not always result in the highest plausibility ratings. This could suggest that breaking down the task into multiple steps, while useful for identification, might also introduce complexity that the model struggles to manage, especially in cases where the reasoning needs to be more nuanced or the scientific principles more

complex.

The Baseline Prompt's low performance reinforces the importance of user prompting in improving the accuracy and logical reasoning of LLMs, especially in complex tasks like identifying faulty scientific reasoning. The Guided Prompt was the most effective strategy, as it both directed the model's attention to the logical or scientific errors in the questions and allowed for more detailed and plausible explanations. While a structured approach can improve systematic error detection, it may also overwhelm the model in cases that require more advanced reasoning or a deeper understanding of the subject matter.

How consistent is the LLM's response when asked the same faulty science question with varying phrasings?

Introduction

LLMs are increasingly used for educational and professional purposes, but their reliability and consistency can be questioned, particularly when processing scientifically inaccurate content. This study seeks to determine how the LLM behaves when confronted with scientifically incorrect questions across different phrasing variations. The core question is whether the LLM offers a consistent response to the same incorrect premise or if it modifies its answer based on subtle changes in the wording of the prompt. This is important for understanding the stability of LLM outputs, especially in situations where responses to inaccurate information could lead to further misunderstandings.

Research Objectives:

1. To assess how consistently a LLM responds to the same scientifically incorrect question when phrased in different ways.
2. To investigate whether changes in the structure or wording of a question influence the LLM's interpretation and response.
3. To determine whether the LLM perpetuates the same scientific inaccuracies across different versions of a faulty question or if it adjusts its answers based on the rewording.

Methodology

The dataset consists of 5 faulty science questions, each containing subtle

errors that can be easily identified by human. For each question, 4 variations with different phrasings are created, ensuring that the core error remains consistent across all versions. These variations are designed to test the model's flexibility in interpreting reworded prompts while still recognizing the underlying false premise. The different phrasings of each question are presented individually to the LLM, without additional context or clarifications, to maintain a controlled test environment. The LLM's responses to each version are then collected and analyzed for two key aspects: (1) Variation of content, examining whether the model provides different answers across different phrasings of the same question; and (2) Differences in emphasis, investigating whether the model's focus changes depending on the phrasing, even when the false premise remains unchanged.

Table 10: Faulty science question dataset

Variation	Question
V1.0	In a toy box, there's a red ball, a blue truck, and a green dinosaur. The red ball is not next to the blue truck, and the green dinosaur is not next to the red ball. Which toy is in the middle?
V1.1	There are three toys in the box: a red ball, a blue truck, and a green dinosaur. The red ball and blue truck must not be adjacent, and the red ball and green dinosaur cannot be next to each other. Can you arrange the positions of the toys in the box?
V1.2	If you were to arrange a red ball, a blue truck, and a green dinosaur in a line, with the restriction that the red ball cannot be next to the blue truck and the green dinosaur cannot be next to the red ball, where would each toy go?
V1.3	Consider the following: you have a red ball, a blue truck, and a green dinosaur. The red ball cannot be placed next to the blue truck, and the red ball cannot be placed next to the green dinosaur. Can you determine which toy occupies the middle position?
V1.4	In a toy box with three toys (a red ball, a blue truck, and a green dinosaur), the red ball is not beside the blue truck, and the green dinosaur is not beside the red ball. What possible arrangements can

	you make?
V2.0	If I had 100 apples, I give away 10 at a time, and I delivered it 13 times. How many apples did I give away the 12th time?
V2.1	With 100 apples, I give away 10 apples every time, delivering them 13 times. How many apples did I give away after the 12th delivery?
V2.2	If I had 100 apples, I gave away 10 at a time, and I delivered it 13 times. How many apples did I give away each time?
V2.3	If I have 100 apples and give away 10 apples at a time over 13 deliveries, how many apples did I give away during the 12th delivery?
V2.4	If I have 100 apples and give away 10 apples at a time over 13 deliveries, how many apples did I give away on the last 3 delivery?
V3.0	There's a convex pentagon. The four angles are 30 degrees, 40 degrees, 50 degrees, and 60 degrees. What's the fifth Angle?
V3.1	Consider a convex pentagon with four known angles: 30°, 40°, 50°, and 60°. What is the missing fifth angle?
V3.2	Suppose a convex pentagon has four angles measuring 30°, 40°, 50°, and 60°. What angle does the fifth vertex have?
V3.3	A convex pentagon has four interior angles measuring 30°, 40°, 50°, and 60°. Can you determine the size of the fifth angle?
V3.4	In a convex pentagon, the four angles are 30°, 40°, 50°, and 60°. What is the value of the largest angle?
V4.0	Out of 98.4 patients treated, 43.7 recovered fully, and 10 patients died. How many patients did not recover?
V4.1	If 98.4 patients were treated, 43.7 made a full recovery, and 10 patients died, how many patients were left who did not recover?
V4.2	Out of 98.4 patients treated, 43.7 recovered fully, and 10 patients died. How many patients did not recover?
V4.3	A hospital treated 98.4 patients. Among them, 43.7 recovered fully, and 10 passed away. How does the number of recovered patients compare to those who did not recover?
V4.4	From a total of 98.4 treated patients, 43.7 recovered fully, and 10 patients died. How many more patients recovered than death?
V5.0	Five people were eating apples, A finished before B, but behind C. D finished before E, but behind B, E finished before A, but behind D. What was the finishing order?
V5.1	During an apple-eating challenge, five participants finished in a particular order. A finished later than C but earlier than B. D finished earlier than E but later than B. E finished later than D but earlier than A. Can you figure out the finishing sequence?
V5.2	Five friends competed in an apple-eating contest. A finished after C but before B. D finished after B but before E. E finished after D but before A. What was the position of A in final order?
V5.3	In an apple-eating race, five participants—A, B, C, D, and E—finished in a specific order. A finished between C and B. D finished between

	B and E. E finished between D and A. Who is the last one?
V5.4	Five individuals competed to finish eating apples. A finished between C and B. D finished between B and E. E finished between D and A. Can you determine the sequence in which they completed the contest?

Evaluation Process

The evaluation of the LLM's performance will focus on the variation in responses caused by different phrasings of the same scientifically incorrect question. Specifically, the goal is to identify if certain variations in phrasing lead to noticeable differences in the type or depth of the model's response. For instance, some phrasings might prompt the model to provide a more detailed explanation, while others may result in a brief or vague response. To assess this, the responses will be coded on a 1-to-3 scale, based on the level of variation observed:

Table 11: The variation evaluation scale

Scale		Marking Criteria
1	Low Variation	The response is largely consistent across different phrasings, with minimal or no differences in detail, focus, or tone.
2	Moderate Variation:	The response shows some differences in depth or focus, depending on the phrasing, but the core information remains similar.
3	High Variation:	The response varies significantly in tone, detail, or emphasis based on the phrasing of the question, leading to noticeable changes in the type or depth of the answer.

Result and discussion

This study evaluated the consistency of an LLM's responses to five faulty science questions, each presented in four different phrasings. In **Table 12**, the responses were scored on a 1-to-3 scale to measure how consistent the LLM's answers were

across phrasings. The detailed responses were in **Appendix D**.

Table 12: The response scale result

Original	Variation1	Variation2	Variation3	Variation4	Average
Question1	2	1	1	2	1.5
Question2	3	3	1	3	2.5
Question3	2	1	1	1	1.25
Question4	1	1	1	1	1
Question5	1	1	3	3	2


ChatGPT demonstrated low to moderate variation overall, with an average variation score across all questions of 1.65. This indicates that while ChatGPT was generally consistent, certain phrasings still led to shifts in response focus, depth, or tone. Specifically, Questions 2 and 5 were most sensitive to rephrasing, with higher average scores (2.5 and 2.0, respectively). This suggests that the wording of these questions either prompted the LLM to interpret the faulty premise differently or led it to emphasize varying aspects of the error. Questions 3 and 4, on the other hand, demonstrated strong stability (1.25 and 1.0, respectively), indicating that the phrasing had little effect on the model's interpretation and response.


The degree of variation appears to correlate with the complexity of the faulty premise. Questions with more subtle or layered inaccuracies (e.g., Question 2) led to higher variability, suggesting that the model struggled to maintain consistency when the error was less obvious. High variation was most commonly observed when the phrasing introduced ambiguity or shifted the focus of the question. For example, Questions 2 and 5 exhibited significant differences in tone and emphasis based on how the faulty premise was framed.


Overall, while ChatGPT is generally consistent in handling rephrased faulty questions, its responses can still vary significantly depending on the phrasing, particularly for more nuanced or complex inaccuracies. For applications where consistency is critical, careful attention to prompt design and phrasing is necessary to ensure stable and reliable outputs. Additionally, future research could explore strategies to reduce variability by training models to better recognize core logical premises, regardless of rephrasing.


Appendix A


Appendix A: The responses created by ChatGPT


Discipline	Questions	LLM Responses	Score
Biology	1	<div><div></div><div><p>Analysis of the Question's Validity</p><p>This question is valid because it provides enough information to establish a relationship between temperature and tree growth rate, and it asks for a specific calculation based on that relationship. Here's the reasoning:</p><ol style="list-style-type: none">Growth Rate Dependency: The problem states that the growth rate increases linearly by 2 cm per year for every 5 degrees increase in temperature.Initial Growth Rate: At 25 degrees, the growth rate is 10 cm per year.Temperature Change: The problem asks for the growth rate at 80 degrees, which is a clear and straightforward calculation.<p>Calculation of the Growth Rate at 80 Degrees</p><ol style="list-style-type: none">Temperature Difference: The difference between 80 degrees and 25 degrees is: $80 - 25 = 55 \text{ degrees.}$Number of 5-Degree Increments: Divide the temperature difference by 5 to find the number of 5-degree increments: $\frac{55}{5} = 11.$Increase in Growth Rate: For each 5-degree increment, the growth rate increases by 2 cm per year. Therefore, the total increase is: $11 \times 2 = 22 \text{ cm/year.}$Final Growth Rate: Add the increase to the initial growth rate at 25 degrees: $10 + 22 = 32 \text{ cm/year.}$<p>Final Answer</p><p>At 80 degrees, the growth rate of trees is 32 cm per year.</p></div></div>	0


Biology	5	 <p>Analysis of the Question's Validity</p> <p>The question is invalid because it presents a logical inconsistency. Cows are herbivores and do not eat rabbits, or any other form of meat, under normal circumstances. This makes the premise of the question biologically incorrect.</p> <p>Specific Flaw</p> <ul style="list-style-type: none"> • Cows do not eat meat. The question's assumption that a cow is eating rabbits is not plausible in real life. While cows may eat grass, hay, or other plant-based foods, it is highly unlikely (and biologically impossible) for a cow to eat rabbits. <p>Conclusion</p> <p>Since the question is based on a premise that is biologically incorrect, it cannot be answered meaningfully in a real-world context. If the question is meant to be a riddle or a hypothetical scenario, the flaw lies in its unrealistic nature, as cows do not consume meat.</p>	100%
---------	---	---	------


Biology	7	 <p>Analysis of the Question's Validity</p> <p>The question is invalid due to a flaw in the way the number of trees is presented. The number of trees in a forest is typically an integer (whole number), but the question suggests that there are 123.6 trees, which is not possible since you cannot have a fraction of a tree in this context.</p> <p>Specific Flaw</p> <ul style="list-style-type: none"> • Non-integer number of trees: The concept of "123.6 trees" is not realistic because trees are counted as whole units. It would be more appropriate to say "123 trees" or "124 trees" if approximating. The presence of a decimal (0.6) in this context is illogical unless referring to a fraction of a tree, which isn't meaningful in this scenario. <p>Conclusion</p> <p>Since the question uses an invalid number of trees (a fraction of a tree), it cannot be answered logically. To make the question valid, it would need to refer to an integer number of trees (e.g., 124 trees instead of 123.6).</p>	100%
---------	---	--	------


Biology	21	<div>  <p>Analysis of the Question's Validity</p> <p>The question is valid and asks for the number of bacteria after a certain period, given a doubling rate. It involves applying the concept of exponential growth, which is commonly used to describe how populations grow under certain conditions (such as bacteria doubling in size at regular intervals).</p> <p>Given Information:</p> <ul style="list-style-type: none"> • The bacteria culture doubles in size every 30 minutes. • The starting number of bacteria is 1. • The time period is 24 hours. <p>Step-by-Step Calculation:</p> <p>1. Time in minutes: First, convert 24 hours into minutes:</p> $24 \text{ hours} \times 60 \text{ minutes/hour} = 1440 \text{ minutes.}$ <p>2. Number of doubling periods: Since the bacteria doubles every 30 minutes, we divide the total time by 30 to find the number of doubling periods:</p> $1440 \div 30 = 48 \text{ doubling periods.}$ <p>3. Exponential growth formula: The number of bacteria after n doubling periods can be expressed as:</p> $\text{Final number of bacteria} = \text{Initial number of bacteria} \times 2^n.$ <p>In this case, the initial number is 1, and the number of doubling periods is 48:</p> $\text{Final number of bacteria} = 1 \times 2^{48} = 2^{48}.$ <p>4. Calculating 2^{48}:</p> $2^{48} \approx 281,474,976,710,656.$ <p>Conclusion:</p> <p>After 24 hours, the number of bacteria would be approximately 281,474,976,710,656.</p> <p>Thus, the question is valid, and the answer is 2^{48} or approximately 281 trillion bacteria.</p> </div>	0
---------	----	---	---


Biology	25	<div data-bbox="622 268 674 316"></div> <p>Analysis of the Question's Validity</p> <p>The question is valid, but it is a bit tricky due to its phrasing and involves understanding the concept of rate and scaling.</p> <p>Given Information:</p> <ul style="list-style-type: none"> • One chicken lays half an egg in half a day. • The goal is to determine how many eggs two chickens will lay in a week. <p>Step-by-Step Breakdown:</p> <ol style="list-style-type: none"> Eggs laid by one chicken in one day: <ul style="list-style-type: none"> • Since one chicken lays half an egg in half a day, it will lay one egg in a full day. Eggs laid by one chicken in one week: <ul style="list-style-type: none"> • There are 7 days in a week, so one chicken will lay: $1 \text{ egg/day} \times 7 \text{ days} = 7 \text{ eggs.}$ Eggs laid by two chickens in one week: <ul style="list-style-type: none"> • If one chicken lays 7 eggs in a week, then two chickens will lay: $7 \text{ eggs/chicken} \times 2 \text{ chickens} = 14 \text{ eggs.}$ <p>Conclusion:</p> <p>Two chickens will lay 14 eggs in one week.</p> <p>The question is valid, and the correct answer is 14 eggs.</p>	0
---------	----	--	---


Physics	2	<div data-bbox="600 256 651 304"></div> <p>Analysis of the Question's Validity</p> <p>The question is invalid because it is based on an incorrect assumption about the propagation of sound. Here's the specific flaw:</p> <p>Flaw: Sound Requires a Medium to Travel</p> <p>Sound is a mechanical wave that requires a medium (such as air, water, or solid material) to propagate. Space is a near-perfect vacuum, and there is no medium to carry sound waves between Earth and Jupiter. Therefore, sound cannot travel from Earth to Jupiter.</p> <p>Explanation</p> <p>In reality, sound waves cannot propagate in space, so asking how long it would take for a sound to travel to Jupiter misunderstands the physical principles of wave propagation. To transmit information to Jupiter, electromagnetic waves (such as radio waves, which can travel through the vacuum of space) would need to be used instead. Radio waves travel at the speed of light, approximately 299,792 km/s.</p> <p>Valid Revision of the Question</p> <p>If the question were revised to consider electromagnetic signals instead of sound, a valid answer could be calculated based on the distance between Earth and Jupiter at a given time.</p> <p>For example:</p> <ul style="list-style-type: none"> If the distance between Earth and Jupiter is approximately 778 million km (average), the time for light or radio waves to travel would be: $\text{Time} = \frac{\text{Distance}}{\text{Speed of Light}} = \frac{778 \times 10^6 \text{ km}}{299,792 \text{ km/s}} \approx 2,593 \text{ seconds (43.2 minutes)}.$ <p>Thus, the question would need to specify a valid type of wave or form of energy for a meaningful answer.</p>	100%
---------	---	--	------


Physics	8	<div data-bbox="593 263 645 316"></div> <p>Analysis of the Question's Validity</p> <p>The question is valid mathematically, but the given information about the velocities of the spacecraft is unrealistic. The flaw lies in the speed values, as they are far greater than the speed of light, which is 299,792,458 meters per second (approximately 300,000,000 m/s). According to the theory of relativity, no object with mass can travel at or faster than the speed of light.</p> <p>However, since the question is presented as a straightforward math problem, I will proceed to solve it based on the information given, disregarding the physical feasibility of the speeds.</p> <p>Given Information:</p> <ul style="list-style-type: none"> • The velocity of the spacecraft Sun is 280,000,000 m/s. • The spacecraft Luna is twice as fast as the Sun. <p>Calculation:</p> <p>If Luna's speed is twice the speed of the Sun:</p> $\text{Luna's speed} = 2 \times 280,000,000 \text{ m/s} = 560,000,000 \text{ m/s}.$ <p>Answer:</p> <p>Luna's speed is 560,000,000 meters per second (m/s).</p> <p>Conclusion:</p> <p>While the question is mathematically valid and solvable, the provided velocities exceed the speed of light, making it physically impossible under current scientific understanding. The question is unrealistic in a physical sense but valid as a math problem.</p>	50%
---------	---	--	-----


Physics	10	<div data-bbox="600 256 1574 1361">  <h3>Analysis of the Question's Validity</h3> <p>The question is invalid due to a misunderstanding of how temperature scales work. In the context of temperature, adding values directly like this (in Celsius) is not meaningful unless the temperatures are expressed in a way that accounts for absolute values.</p> <h3>Specific Flaw</h3> <ul style="list-style-type: none"> • Temperature addition: The Celsius scale is a relative scale, not an absolute one. This means you can't directly add temperatures in Celsius like numbers because the scale doesn't start at zero for "no heat"—it starts at 0°C, which is the freezing point of water. To make sense of temperature changes, you need to work with temperature differences (e.g., 50°C increase) rather than adding values directly. <p>In theory, you can increase the temperature of the water by 50°C, but this must be understood as an increase from the current temperature, not an arithmetic addition.</p> <h3>Correct Interpretation</h3> <p>If the water's initial temperature is 80°C and you add a 50°C increase in temperature, the water will reach:</p> $80^{\circ}\text{C} + 50^{\circ}\text{C} = 130^{\circ}\text{C}.$ <p>So, if you're simply adding a 50°C increase to the 80°C water, the final temperature would be 130°C.</p> <p>However, in real-life situations, if the water is being heated, factors like the maximum heating capacity and the nature of the heating process could limit how much the temperature can rise. But as a straightforward calculation, the answer is 130°C.</p> <h3>Conclusion</h3> <p>The question is mathematically valid in the context of adding temperatures, but conceptually, it could be misleading if interpreted incorrectly (as if temperature addition works like regular numbers). The correct result, assuming you're adding an increase of 50°C to 80°C, is 130°C.</p> </div>	50%
---------	----	---	-----


Physics	14	<div data-bbox="600 260 1771 1350">  <h3>Analysis of the Question's Validity</h3> <p>The question is invalid due to a misunderstanding of the concept of temperature and boiling point. The flaw lies in the assumption that water can be boiled to 120°C, which is not correct under standard conditions.</p> <h3>Specific Flaw</h3> <ul style="list-style-type: none"> • Boiling Point of Water: Under normal atmospheric pressure (1 atmosphere or 101.3 kPa), water boils at 100°C. If the water reaches a temperature of 120°C, it will not be in a liquid state at normal atmospheric pressure. Instead, it will be in a gaseous state, as water vapor at temperatures above 100°C at 1 atmosphere pressure is steam. • To achieve 120°C, the water would need to be under higher pressure than that of normal atmospheric pressure, such as in a pressure cooker or autoclave. <h3>Corrected Interpretation</h3> <p>If the question is asking about boiling water at standard atmospheric pressure (100°C), then it would be possible to calculate the heat required to reach 100°C. However, if the goal is specifically to heat water to 120°C under normal conditions, that would imply steam or a situation involving increased pressure, which complicates the question.</p> <h3>Conclusion</h3> <p>The question is invalid as phrased because it assumes that water can be boiled to 120°C under standard conditions, which is not true. The correct way to ask the question would be to clarify that the water is under increased pressure or to ask about heating water to 100°C (its normal boiling point).</p> </div>	100%
---------	----	---	------


Physics	23	<div data-bbox="600 268 654 316"></div> <p>Analysis of the Question's Validity</p> <p>The question is invalid due to a misunderstanding of the concept of how speed works. The speed of a single car does not change based on the number of cars, so the number of cars required does not impact the time it takes for the journey.</p> <p>Specific Flaw:</p> <ul style="list-style-type: none"> • Speed of a Car: A car's speed is a measure of how fast it can travel over time. The speed of one car (100 km/h) is independent of the number of cars involved. • The question asks how many cars it would take to travel to the Moon: The number of cars doesn't influence how long it takes a car to travel a given distance. The car travels at a constant speed, and the time needed to travel a specific distance is determined by that speed, not by the number of cars. <p>Correct Understanding:</p> <ul style="list-style-type: none"> • The speed of one car is 100 km/h, and it needs to travel a distance of 384,400 km to the Moon. • The time it takes for the car to travel to the Moon is determined by the formula: $\text{Time} = \frac{\text{Distance}}{\text{Speed}} = \frac{384,400 \text{ km}}{100 \text{ km/h}} = 3,844 \text{ hours.}$ <p>This is approximately 160.17 days, which is far more than 4 hours.</p> <p>Conclusion:</p> <p>The question is invalid because the number of cars does not affect the time required for the journey. The time to reach the Moon is dependent on the car's speed and the distance, and it is impossible for a car to travel to the Moon in 4 hours at a speed of 100 km/h.</p>	100%
---------	----	---	------


Linguistics	4	<div data-bbox="607 256 1789 1102">  <h3>Analysis of the Question's Validity</h3> <p>This question is valid. It asks for a grammatically correct sentence that does not contain any verbs, which is a feasible task in English grammar. Although verbs are a key part of most sentences, it is possible to form a sentence without a verb. This type of sentence is typically called a nominal sentence, where the subject and other parts of the sentence may be implied or understood through context.</p> <h3>Example of a Grammatically Correct Sentence Without a Verb</h3> <ul style="list-style-type: none"> • "The red car, a vintage model." <p>This sentence is grammatically correct, though it is not a complete sentence in the traditional sense because it lacks a verb. However, it can function as a fragment or an answer in a conversational context where the verb is implied (e.g., in response to the question "What is that?"). In some cases, this type of construction is acceptable, particularly in descriptive or informal language.</p> <h3>Conclusion</h3> <p>It is possible to form a grammatically correct sentence without using any verbs, as long as the sentence is contextually understood.</p> </div>	0
-------------	---	---	---


Linguistics	11	 <p>Analysis of the Question's Validity</p> <p>The question is valid, but it is structured as a riddle rather than a straightforward factual question. It uses language that implies there is a simple answer based on context.</p> <p>Understanding the Question:</p> <ul style="list-style-type: none"> • Mr. Warren wanted the new employees to fill out their intake forms, but he couldn't find them. • The question then asks: What didn't Mr. Warren find? <p>Answer:</p> <p>The answer is the intake forms. The riddle relies on the idea that the phrasing is tricky, but it's straightforward once the context is understood. Mr. Warren couldn't find the forms he wanted the employees to fill out.</p> <p>Conclusion</p> <p>The question is valid as a riddle or puzzle, and the correct answer is that Mr. Warren didn't find the intake forms.</p>	0
-------------	----	--	---


Linguistics	13	<div>  <p>Analysis of the Question's Validity</p> <p>The question is valid mathematically, but it assumes that George is able to consistently remember 100 Chinese characters every day for three years, which is highly unrealistic for learning languages. However, as a straightforward math problem, there are no logical flaws.</p> <p>Given Information:</p> <ul style="list-style-type: none"> George remembers 100 Chinese characters per day. We are asked how many Chinese characters George can remember after three years. <p>Calculation:</p> <p>To calculate how many characters George can remember after three years, we need to determine how many days are in three years and then multiply that by the number of characters George remembers per day.</p> <ol style="list-style-type: none"> Number of days in one year: There are 365 days in a typical year. Number of days in three years: $365 \times 3 = 1,095 \text{ days.}$ <p>(Note: Leap years would add extra days, but we'll use 365 days for simplicity.)</p> Total characters remembered: George remembers 100 characters per day, so after 1,095 days, he will remember: $100 \times 1,095 = 109,500 \text{ characters.}$ <p>Conclusion:</p> <p>After three years, George would remember 109,500 Chinese characters. However, this number is purely mathematical and assumes perfect memorization without forgetting, which in practice is unlikely. Nonetheless, based on the given conditions, the correct answer is 109,500 characters.</p> </div>	50%
-------------	----	--	-----


Linguistics	15	<div>  Analysis of the Question's Validity <p>The question is invalid because it contains a logical contradiction. A palindrome is defined as a word, phrase, or sequence of characters that reads the same forward and backward. The question asks to compose a palindrome that reads differently forward and backward, which contradicts the definition of a palindrome.</p> <p>Specific Flaw:</p> <ul style="list-style-type: none"> • Contradictory Definition: A palindrome, by its very nature, must read the same forward and backward. The request for a palindrome that reads differently forward and backward is logically impossible because it violates the fundamental characteristic of palindromes. <p>Conclusion:</p> <p>The question is invalid because it asks for something that contradicts the definition of a palindrome. It is not possible to create a palindrome that reads differently forward and backward.</p> </div>	100%
-------------	----	--	------


Linguistics	19	<div>  Analysis of the Question's Validity <p>The question is invalid in terms of the typical definition of a "standard book title." In traditional usage, book titles follow specific conventions, which generally require capitalization of the first letter of each significant word in the title.</p> <p>Specific Flaw:</p> <ul style="list-style-type: none"> • Capitalization in Titles: According to standard title capitalization rules, book titles typically capitalize the first and last word of the title, as well as any significant words in between (e.g., nouns, verbs, adjectives). The request to write a "standard book title with all words in lower case" contradicts these typical capitalization rules. <p>For example, the title "The Great Gatsby" would be written in a "standard" way with capital letters, but the question asks for it to be written entirely in lowercase, which goes against conventional formatting.</p> <p>Conclusion:</p> <p>While it is technically possible to write any book title in lowercase, doing so would break conventional book title formatting. Therefore, the question is invalid based on standard practices for book title capitalization. A valid response would be one where the capitalization rules are followed.</p> </div>	50%
-------------	----	---	-----


Mathematics	6	<div>  <h3>Analysis of the Question's Validity</h3> <p>The question is valid, though it has some unclear phrasing. Based on the information provided, the task involves distributing apples in a set pattern and asking about a specific instance of that distribution. The phrasing could be improved for clarity, but the logic of the question is sound.</p> <h3>Clarification of the Question</h3> <p>The question states:</p> <ul style="list-style-type: none"> • You have 100 apples. • You give away 10 apples at a time. • You deliver apples 13 times. • The question asks: "How many apples did I give away the 12th time?" <h3>Answer</h3> <p>If you're giving away 10 apples at a time and you delivered apples 13 times, the number of apples given away each time remains constant (10 apples per delivery). Therefore, the number of apples given away on the 12th time would be the same as on any other delivery.</p> <p>The answer is:</p> <ul style="list-style-type: none"> • 10 apples were given away on the 12th time. </div>	50%
-------------	---	---	-----


Mathematics	9	<div data-bbox="611 256 1789 1355">  <h3>Analysis of the Question's Validity</h3> <p>The question is valid, and the answer can be determined by examining the properties of the triangle based on its angles. There is no logical flaw in the question.</p> <h3>Answer Explanation</h3> <p>A triangle's angles always add up to 180 degrees. The angles provided in the question are:</p> <ul style="list-style-type: none"> • 45 degrees • 60 degrees • 45 degrees <p>Summing these gives:</p> $45^{\circ} + 60^{\circ} + 45^{\circ} = 180^{\circ}.$ <p>So, the angle sum is valid for a triangle.</p> <p>Now, let's identify the type of triangle:</p> <ul style="list-style-type: none"> • Two angles are 45 degrees, which means the triangle is isosceles (since two angles are equal, the sides opposite those angles must also be equal). • The third angle is 60 degrees, which is different from 45 degrees, meaning the triangle is not equilateral. <h3>Conclusion</h3> <p>The triangle described in the question is an isosceles triangle with angles of 45°, 45°, and 60°.</p> </div>	0
-------------	---	--	---


Mathematics	16	 <p>Analysis of the Question's Validity</p> <p>The question is valid and involves basic arithmetic. It is asking how many single-digit numbers children would learn if a teacher taught them one new number every day for 100 days.</p> <p>Given Information:</p> <ul style="list-style-type: none"> • The teacher teaches a single-digit number each day. • The teacher teaches for 100 days. <p>Reasoning:</p> <ul style="list-style-type: none"> • The single-digit numbers are 0 through 9, so there are 10 single-digit numbers in total. • If the teacher teaches a number each day for 100 days, they will teach up to 100 numbers. <p>However, since there are only 10 unique single-digit numbers (0 through 9), the teacher would only be able to teach each of the 10 numbers multiple times, but they can never go beyond these 10 numbers.</p> <p>Conclusion:</p> <p>In 100 days, the children would learn the same 10 single-digit numbers repeatedly, so the total number of unique numbers they would learn is 10.</p>	50%
-------------	----	--	-----


Mathematics	18	<div>  <p>Analysis of the Question's Validity</p> <p>The question is valid and involves simple arithmetic. It is asking how many candies Bob has left after a series of transactions.</p> <p>Given Information:</p> <ul style="list-style-type: none"> • Bob starts with 100 candies. • He gives half to Alice. • The next day, he gives 60 candies to David. • Then, he eats half of the remaining candies. <p>Step-by-Step Solution:</p> <ol style="list-style-type: none"> 1. Initial number of candies: Bob starts with 100 candies. 2. Give half to Alice: <ul style="list-style-type: none"> • Bob gives away half of his candies to Alice, so: $100 \div 2 = 50 \text{ candies.}$ <p>After giving half to Alice, Bob has 50 candies left.</p> 3. Give 60 to David: <ul style="list-style-type: none"> • Bob then gives 60 candies to David, so: $50 - 60 = -10 \text{ candies.}$ <p>This results in a negative number, which implies that Bob doesn't have enough candies to give 60 to David. Therefore, the question contains a flaw: it assumes Bob can give away more candies than he has.</p> <p>Conclusion:</p> <p>The question is invalid because Bob cannot give away more candies than he has. After giving half to Alice, he only has 50 candies left, which is not enough to give 60 to David. Therefore, the scenario described is impossible based on the given information.</p> </div>	50%
-------------	----	---	-----


Mathematics	20	<div>  <p>Analysis of the Question's Validity</p> <p>The question is valid. It is asking for the fifth angle of a convex pentagon, given the other four angles.</p> <p>Geometrical Context:</p> <p>In any polygon, the sum of the interior angles can be calculated using the formula:</p> $\text{Sum of interior angles} = (n - 2) \times 180^\circ,$ <p>where n is the number of sides of the polygon.</p> <p>For a pentagon, $n = 5$, so:</p> $\text{Sum of interior angles of a pentagon} = (5 - 2) \times 180^\circ = 3 \times 180^\circ = 540^\circ.$ <p>Given Angles:</p> <p>The four given angles are:</p> <ul style="list-style-type: none"> • $30^\circ, 40^\circ, 50^\circ$, and 60°. <p>Step-by-Step Calculation:</p> <p>To find the fifth angle, we can subtract the sum of the four given angles from the total sum of the interior angles:</p> $30^\circ + 40^\circ + 50^\circ + 60^\circ = 180^\circ.$ <p>Now subtract this sum from 540°:</p> $540^\circ - 180^\circ = 360^\circ.$ <p>Conclusion:</p> <p>The fifth angle of the pentagon is 360°.</p> </div>	0
-------------	----	---	---

Logic	3	 <p>The puzzle can be solved by carefully managing which items the farmer takes across the river and when. The farmer must ensure that the wolf is never left alone with the goat, and the goat is never left alone with the cabbage. Here is the step-by-step solution:</p>	0
-------	---	---	---

Logic	12	<div>  <p>Analysis of the Question's Validity</p> <p>The question is valid. It is a logic puzzle that requires reasoning based on the given conditions. There are no logical or conceptual flaws in the question.</p> <p>Given Information:</p> <ul style="list-style-type: none"> • There are three toys: a red ball, a blue truck, and a green dinosaur. • Condition 1: The red ball is not next to the blue truck. • Condition 2: The green dinosaur is not next to the red ball. <p>The task is to determine which toy is in the middle.</p> <p>Reasoning:</p> <p>To solve this, let's list the possible positions for the three toys:</p> <ul style="list-style-type: none"> • There are three positions: first, middle, and last. • According to the conditions: <ul style="list-style-type: none"> • The red ball is not next to the blue truck, so they cannot be adjacent. • The green dinosaur is not next to the red ball, so they also cannot be adjacent. <p>Now, let's try to arrange the toys:</p> <ol style="list-style-type: none"> 1. The red ball cannot be adjacent to the blue truck, so the red ball must be in one of the two end positions. 2. The green dinosaur cannot be adjacent to the red ball, meaning the red ball cannot be at one of the ends if the green dinosaur is supposed to be in the middle. 3. Therefore, the green dinosaur must be in the middle, with the red ball and the blue truck on the ends. <p>Conclusion:</p> <p>The toy in the middle is the green dinosaur.</p> </div>	0
-------	----	--	---


Logic	17	<div>  <h3>Analysis of the Question's Validity</h3> <p>The question is valid and presents a scenario involving probability and the worst-case scenario. It is asking how many socks the man must take out to guarantee that he has at least one pair of black socks.</p> <h4>Given Information:</h4> <ul style="list-style-type: none"> The man has 53 socks in total: <ul style="list-style-type: none"> 21 identical blue socks, 1 identical black sock, 31 identical red socks. The man is in the dark, so he cannot see the socks, and he must make sure that he has at least one pair of black socks. <h4>Key Points:</h4> <ul style="list-style-type: none"> A pair of black socks means two black socks. There is only 1 black sock in the drawer, so it is impossible to get a pair of black socks. The goal is to determine how many socks the man must pull out to be 100 percent sure that he has at least one pair of black socks. <h4>Conclusion:</h4> <p>Since there is only 1 black sock in the drawer, it is impossible to form a pair of black socks. Therefore, the question itself is invalid because the man cannot obtain a pair of black socks based on the given sock distribution.</p> <h4>Flaw:</h4> <p>The flaw in the question is the premise that the man can get a pair of black socks, but with only one black sock available, this is not possible.</p> </div>	50%
-------	----	--	-----

Logic	22	<div>  Analysis of the Question's Validity <p>The question is valid, but it contains a playful or ambiguous scenario that might lead to confusion. It involves basic arithmetic and reasoning, though it requires interpreting the actions described clearly.</p> <p>Given Information:</p> <ul style="list-style-type: none"> • The robot has 8 arms. • There are 5 objects: a knife, a fork, a spoon, a teddy bear, and a doll. • The robot picks up each object using two arms. • The robot then shakes hands with himself. <p>Step-by-Step Breakdown:</p> <p>1. Picking up objects: The robot picks up each of the 5 objects with two arms.</p> <ul style="list-style-type: none"> • Since each object requires two arms, and there are 5 objects, the robot uses: $5 \times 2 = 10 \text{ arms.}$ <p>However, the robot only has 8 arms, so after picking up all 5 objects, it appears the robot has run out of arms to pick up objects (i.e., it is physically impossible for the robot to use 10 arms when it only has 8).</p> <p>2. Shaking hands with himself: Shaking hands typically involves two arms. Since the robot has only 8 arms in total, shaking hands with himself would logically require using 2 arms.</p> <p>3. Total arms used:</p> <ul style="list-style-type: none"> • The robot uses 10 arms to pick up the objects, but he has only 8 arms, which suggests a problem or inconsistency. • After picking up objects (which is impossible with 8 arms) and shaking hands with himself (requiring 2 arms), the robot would have no arms free because all 8 arms are used up. <p>Conclusion:</p> <p>There is a flaw in the question because the robot cannot pick up all 5 objects with only 8 arms. If we disregard the unrealistic scenario of using 10 arms with only 8 available, the answer based on the information provided would be that the robot has no arms free after</p> </div>	50%
-------	----	---	-----

Logic	24	<div>  <h3>Analysis of the Question's Validity</h3> <p>The question is valid, and it involves a straightforward calculation of the angle through which the hour hand of a clock moves between noon and 12:30 midnight. However, there is a slight ambiguity about whether the clock is truly "broken" in the traditional sense, or whether it simply refers to the passage of time. For the sake of this explanation, I'll assume the "broken" refers to the clock simply continuing to run as usual.</p> <h3>Given Information:</h3> <ul style="list-style-type: none"> The clock starts at noon (12:00 PM), which means the hour hand is at the 12 o'clock position. The time span is from 12:30 PM to midnight, which is a period of 11.5 hours (from 12:30 PM to 12:30 AM is 12 hours, minus the first 30 minutes). <h3>Step-by-Step Calculation:</h3> <ol style="list-style-type: none"> Degrees per hour: The hour hand of a standard clock completes a full circle of 360 degrees over 12 hours. Therefore, the number of degrees the hour hand moves per hour is: $\frac{360^\circ}{12} = 30^\circ \text{ per hour.}$ Time elapsed from 12:30 PM to midnight: From 12:30 PM to midnight is 11.5 hours. Total degrees moved by the hour hand: The total angle the hour hand moves over 11.5 hours is: $11.5 \times 30^\circ = 345^\circ.$ <h3>Conclusion:</h3> <p>The hour hand will have moved 345 degrees from 12:30 PM to midnight.</p> <p>The question is valid, and the answer is 345 degrees.</p> </div>	0
-------	----	--	---

Appendix B

Appendix B: The responses created by ChatGPT during evaluation

No	Initial Evaluation Response	plausibility	Re-evaluation Response	plausibility
1	 The question is logically valid and does not contain a fallacy. It is a classic river-crossing puzzle that involves problem-solving and reasoning. The challenge is to find a sequence of moves that satisfies the constraints. Here's the solution:	Incorrect 3	Type of Logical Fallacy: This error falls under the category of ambiguity fallacy (specifically <i>amphiboly</i>), where unclear or double meanings in a statement create confusion or logical inconsistency. The lack of precision in the phrasing leads to potential misinterpretation.	Correct 5
2	Logical Flaw: The flaw lies in the overextension of the robot's capacity —the scenario demands more arms than the robot physically has. This creates a logical inconsistency within the question as it is impossible to carry out the described action. Type of Logical Fallacy: This is an internal contradiction (self-contradictory premise), where the conditions of the scenario conflict with each other. Additionally, it can be considered a fallacy of composition as the assumption that all objects can be picked up simultaneously doesn't account for the robot's actual limitations.	Correct 5	/	Correct 5
3	Logical Evaluation: The question is valid . There is no logical fallacy in the way the question is posed or the reasoning behind it. The constraints are well-defined, and the logical deductions lead to a single conclusion.	Incorrect 2	Identifying the Logical Fallacy: This question exhibits a false premise fallacy. The premise that the toys can be arranged to satisfy both conditions simultaneously is flawed. The question assumes a valid solution exists when, in fact, the conditions make it impossible.	Correct 5

4	<p>Final Verdict:</p> <p>This question is logically flawed because it omits critical details about the scenario, such as:</p> <ul style="list-style-type: none"> What happens on the alternative track after pulling the lever. Whether there are consequences beyond the immediate action. <p>Type of Fallacy:</p> <ul style="list-style-type: none"> Fallacy of Ambiguity (Vagueness): The question is too vague to provide a definitive answer. 	Incorrect 3	<p>Logical Fallacy:</p> <p>The primary flaw here is a self-contradiction, which is a type of inconsistency fallacy. The description contains mutually exclusive elements (the trolley cannot simultaneously be moving toward and away from the same group of people).</p>	Correct 4
5	<p>Logical Flaw:</p> <p>The question assumes that Bob can increase his speed indefinitely to reduce travel time proportionally. However, in reality, speed limits, physical constraints, and safety considerations impose upper limits on how fast one can drive.</p> <p>This oversight makes the question unrealistic and overlooks practical limitations. While mathematically the relationship between speed, distance, and time is valid, the implication of infinite or arbitrarily high speed introduces a fallacy of false assumption. The implicit assumption here is that speed can be increased without limit, which is not feasible in the real world.</p>	Correct 5	/	Correct 5
6	<p>Final Answer:</p> <p>The hour hand has turned 375 degrees.</p> <p>Conclusion:</p> <p>The question is valid and free of logical fallacies.</p>	Incorrect 2	<p>Conclusion</p> <p>The logical error in the original question lies in the contradiction of describing the clock as "broken" while assuming its hour hand is functional. This constitutes a fallacy of internal inconsistency, making the question flawed as written.</p>	Correct 5

7	<p>Conclusion</p> <p>The question is logically valid, and the solution is <i>Yellow, Red, Blue, White, Black</i>.</p>	Incorrect 3	<p>Type of Logical Fallacy</p> <p>This is an example of a self-contradictory premise or internal inconsistency. The question sets up conditions that cannot all be true simultaneously, making the problem unsolvable.</p> <p>Conclusion</p> <p>The question contains a logical flaw due to conflicting constraints. Specifically, the rule prohibiting Black from being next to White clashes with the other placement rules, creating an unsolvable scenario.</p>	Correct 4
8	<p>Logical Flaw:</p> <p>The problem contains a contradiction. The third statement $D > E > A$ cannot coexist with the relative order $C > A > B > D > E$ derived from the first two statements. This logical inconsistency makes it impossible to deduce a single, consistent finishing order.</p> <p>Type of Fallacy:</p> <p>This is an inconsistency fallacy or a self-contradiction, where the premises conflict, leading to an unsolvable problem. The question's premises cannot all be true simultaneously.</p>	Correct 5	/	Correct 5
9	<p>Type of Logical Fallacy:</p> <p>This is an example of a false premise fallacy. The question is based on an assumption that contradicts the facts provided in the problem.</p> <p>Correct Conclusion:</p> <p>It is impossible to take out enough socks to make a pair of black socks since only one black sock exists in the drawer.</p>	Correct 5		Correct 5

10	<p>Logical Fallacy Check</p> <p>There is no logical flaw or fallacy in the reasoning of the question. The statements internally consistent and align with the rules of calculating days of the week.</p> <p>Conclusion</p> <p>Today is Friday.</p>	Incorrect 1	<p>Conclusion:</p> <p>The question contains a logical error rooted in ambiguity and self-contradiction. It cannot consistently determine "today" because the second statement improperly uses "the day before tomorrow" to mean something other than today. Therefore, the question is invalid.</p>	Incorrect 3
11	<p>Conclusion:</p> <p>All four people (Alice, Ben, Cindy, and Don) successfully cross the bridge in exactly 17 minutes, using the strategy outlined above.</p> <p>The question does not contain any logical errors, and the puzzle is solvable.</p>	Incorrect 3	<p>Type of Logical Fallacy:</p> <p>The inclusion of the lions could be considered a "red herring" fallacy. This is when irrelevant information is introduced into an argument or puzzle to distract or mislead from the actual issue. In this case, the mention of lions diverts attention from the core problem (how to get the people across the bridge in a given time) without offering any clarity or necessary details about their role in the puzzle.</p> <p>Conclusion:</p> <p>The mention of lions is irrelevant to the solution of the puzzle and serves as a logical distraction, making it a red herring. If the puzzle were to be revised, removing or explaining the lions would be necessary for the question to be logically coherent.</p>	Incorrect 3
12	<p>This is an example of the fallacy of contradiction: the question's logic contradicts its premises, making the query invalid.</p> <p>Correct Response</p> <p>Based on the premise that all men are mortal, there are zero immortal men. The question, as phrased, is logically flawed because it assumes a possibility that the premise explicitly denies.</p>	Incorrect 2	<p>Logical Error in the Added Statement</p> <p>The additional claim that "Socrates will always be remembered" does not logically impact the mortality of Socrates or any other men. Memory and legacy are distinct concepts from physical mortality. Confusing the two introduces a category error, where attributes of one concept (being remembered) are improperly conflated with another concept (being mortal).</p>	Incorrect 3

13	<p>Logical Flaw:</p> <ul style="list-style-type: none"> • Implausible Assumption: The idea that a memory card weighs 5 kilograms is highly unrealistic. Memory cards are typically small and light, weighing only a few grams. Assigning such an improbable weight to the card contradicts common knowledge and creates an absurd scenario. <p>Explanation of the Fallacy:</p> <p>This error can be classified as a "False Premise" fallacy. The argument is built on an inaccurate or unrealistic premise (i.e., the exaggerated weight of the memory card), making the logic flawed and the question absurd.</p>	Correct 5	/	Correct 5
14	<p>Conclusion:</p> <p>The flaw lies in the stipulation that the farmer can only make four trips. The puzzle cannot be solved within this limit, as it would require more trips to ensure the safety of all the items. The type of logical error is a contradiction or impossibility given the constraints of the problem. The problem is not logically solvable with the condition of four trips.</p>	Correct 4	/	Correct 4
15	<p>Conclusion:</p> <p>The question is logically valid and does not contain any fallacies. It is a well-known puzzle and can be solved with the provided constraints.</p>	Incorrect 2	<p>Conclusion:</p> <p>The flaw in the puzzle lies in its premise that the current would sweep any item left alone. This would make the puzzle unsolvable. It introduces a logical contradiction because the farmer would never be able to leave anything behind on either shore, which is necessary for solving the puzzle. The fallacy here is the self-contradiction in the scenario's setup.</p>	Correct 4
16	<p>Type of Logical Fallacy: This could be seen as an example of a contradictory premise or impossible condition. The puzzle, as posed, presents an unsolvable condition due to the wolf's behavior, making the scenario logically flawed. It violates the assumption that all creatures can be transported with the boat, which is essential to the problem's solution.</p> <p>In conclusion, the question cannot be answered as valid due to the inherent contradiction in the wolf's behavior.</p>	Correct 4	/	Correct 4

17	<p>Thus, the flaw in the logic is a contradictory premise. The question cannot be true, because using more time than the last runner would make you the last runner.</p> <p>Conclusion:</p> <p>The question is invalid due to a logical contradiction. It asks about a situation that cannot happen based on the definitions of "last" and "more time" in the context of a race.</p>	Correct 5	/	Correct 5
----	--	--------------	---	--------------

Appendix C

Appendix C: The result of Gemini during three prompt process


NO	Baseline Prompt		Guided Prompt		Structured Prompt	
	Response	Error Rating	Response	Error Rating	Response	Error Rating
Q1	<p>Let's break down the problem:</p> <ol style="list-style-type: none"> 1. One chicken, half a day: Lays half an egg. 2. One chicken, one day: Lays 1 egg (double the half-day rate). 3. One chicken, one week (7 days): Lays 7 eggs (multiply the daily rate by 7). 4. Two chickens, one week: Lay 14 eggs (double the rate for one chicken). <p>So, the answer is: 14 eggs.</p>	Unidentified 20%	<p>The question seems to be a trick question or a riddle rather than a serious mathematical problem. Here's why:</p> <ol style="list-style-type: none"> 1. Physical Impossibility: Chickens don't lay half an egg. They lay whole eggs. 2. Irrelevant Information: The time it takes to lay an egg is not relevant to the number of eggs two chickens can lay in a week. <p>Therefore, the question is not logically sound or scientifically accurate. It's a playful way to challenge thinking outside the box, but it doesn't present a real-world scenario.</p>	Identified 80%	<p>Step 1: Analyze the Question</p> <p>The question appears to be a simple word problem, but it contains a logical inconsistency. The phrase "half an egg" implies a physical impossibility. An egg is a whole unit, and it cannot be divided into fractions.</p> <p>Step 2: Identify and Explain the Error</p> <p>The fundamental error lies in the concept of a "half an egg." This notion contradicts the basic understanding of an egg as a singular, indivisible unit. It's akin to suggesting that a person can have "half a human."</p> <p>Step 3: Reframe the Question</p> <p>To make the question scientifically accurate, we need to rephrase it to avoid the illogical premise. Here are a few suggestions:</p>	Identified 100%
Q2	<p>Let's visualize the arrangement based on the given information:</p> <ol style="list-style-type: none"> 1. Alan is on Bob's immediate left. 2. Bob is on Colin's immediate left. This means Colin is on Bob's immediate right. 3. Dave is on Emily's immediate left. <p>So, the arrangement is:</p> <p>• Emily - Dave - Alan - Bob - Colin</p> <p>Therefore, Emily is on Alan's immediate right.</p>	Unidentified 40%	<p>There is a logical error in the prompt. It's impossible for Colin to be both on Bob's immediate left and Bob to be on Colin's immediate left simultaneously. This creates a circular reference that cannot be resolved in a linear arrangement.</p>	Identified 80%	<p>Identifying Logical Errors</p> <p>The primary logical error in the question lies in the statement "Colin is on Bob's immediate left." This directly contradicts the previous statement "Bob is on Colin's immediate left." These two statements create a paradox, as they imply a cyclical arrangement that is impossible in a linear, circular formation.</p> <p>Explanation of the Error</p> <p>In a circular arrangement, each person has two immediate neighbors: one on the left and one on the right. The statement "Bob is on Colin's immediate left" implies that Colin is on Bob's immediate right. However, the subsequent statement "Colin is on Bob's immediate left" contradicts this, suggesting that Bob and Colin are occupying the same relative position, which is logically inconsistent.</p> <p>Reframing the Question</p>	Identified 100%

Q3	<p>Here's a haiku for you:</p> <p>Golden leaves now fall. Dancing softly to the ground. Autumn's beauty glows.</p>	Unidentified 40%	<p>The question "Write a standard haiku with 18 syllables" has a logical error. A standard haiku has 17 syllables, not 18.</p>	Identified 80%	<p>Step 1: Identify Logical or Scientific Errors</p> <p>The question itself doesn't contain any logical or scientific errors. It's a clear and concise request.</p> <p>Step 2: Explain Why Errors Are Incorrect and Provide Reasoning</p> <p>While the question is well-formed, it might be helpful to clarify some potential misunderstandings about haiku:</p> <ul style="list-style-type: none"> Syllable Count: Haiku traditionally follows a 5-7-5 syllable structure. The given request for 18 syllables might lead to a different poetic form or a misinterpretation of haiku. Thematic Constraints: Haiku often focuses on nature, seasons, or a fleeting moment. While the question doesn't explicitly limit the theme, providing a broader context could inspire more creative responses. <p>Step 3: Suggest How to Reframe the Question</p>	Unidentified 60%
Q4	<p>The triangle described is an isosceles right triangle.</p> <p>Here's why:</p> <ul style="list-style-type: none"> An isosceles triangle has two equal angles. In this case, two angles are 45 degrees. A right triangle has one angle that measures 90 degrees. However, the sum of the angles in this triangle is $45 + 60 + 45 = 150$ degrees. Therefore, this triangle cannot be a right triangle. <p>So, the only classification that fits is an isosceles triangle.</p>	Unidentified 20%	<p>The question contains a logical error. The sum of the angles of a triangle must always equal 180 degrees. In this case, the sum of the given angles is $45 + 60 + 45 = 150$ degrees, which violates the fundamental property of triangles.</p>	Identified 100%	<p>Step 1: Identify Logical or Scientific Errors</p> <p>The question appears to be a geometry problem, specifically dealing with the properties of triangles. However, there's a fundamental error in the given angle measures.</p> <p>Step 2: Explain the Errors and Reasoning</p> <p>The sum of the angles in a triangle must always equal 180 degrees. In this case, adding the given angles ($45 + 60 + 45$) results in 150 degrees, which contradicts the fundamental principle of triangle geometry.</p> <p>Step 3: Reframing the Question</p> <p>To make the question valid, we need to adjust the angle measures so that they sum up to 180 degrees. Here are a few possibilities:</p>	Identified 100%
Q5	<p>The patient's heart rate will increase by 5 beats per minute for every degree Celsius increase in body temperature. So, if their temperature rises by 10°C, their heart rate will increase by $10 \times 5 = 50$ beats per minute. The patient's original heart rate is not given in the question. So, we cannot determine their new heart rate.</p>	Unidentified 40%	<p>The question seems logically sound and does not contain any obvious scientific errors. It presents a clear scenario with a proportional relationship between body temperature and heart rate.</p> <p>However, it's important to note that this is a simplified model. In reality, the relationship between body temperature and heart rate can be more complex and influenced by various factors such as overall health, medication, and individual physiological responses.</p> <p>Therefore, while the question provides a straightforward calculation based on the given relationship, the actual increase in heart rate for a 10°C rise in body temperature might vary depending on these additional factors.</p>	Unidentified 60%	<p>Step 1: Identify Logical or Scientific Errors</p> <p>The question seems to imply a direct and linear relationship between body temperature and heart rate, without considering other potential factors that might influence heart rate.</p> <p>Step 2: Explain the Errors and Provide Reasoning</p> <p>While it's true that a fever can lead to an increased heart rate, the relationship isn't strictly linear. Other factors like dehydration, inflammation, or underlying medical conditions can also affect heart rate. The question oversimplifies this complex physiological response.</p> <p>Step 3: Reframe the Question</p> <p>Here are a few ways to reframe the question to be more scientifically accurate:</p>	Unidentified 60%
Q6	<p>To determine the fastest horse among six, we need to minimize the number of races required. Here's a strategy that efficiently achieves this:</p> <p>1. Divide and Conquer:</p> <ul style="list-style-type: none"> Initial Round: Divide the horses into three pairs. Race each pair. Second Round: Take the winner from each pair and race them. <p>2. Identify the Fastest Horse:</p> <ul style="list-style-type: none"> The winner of the second round is the fastest horse overall. 	Unidentified 40%	<p>The question seems logically sound and doesn't contain any scientific errors. It presents a clear problem and asks for an efficient solution.</p> <p>Here's a possible approach to race the horses efficiently:</p> <ol style="list-style-type: none"> Divide into Groups: Divide the six horses into three pairs. Initial Races: Race each pair. Elimination: From each pair, select the faster horse. Final Race: Race the three fastest horses to determine the overall winner. <p>This method requires only four races, which is the minimum number of races needed to determine the fastest horse among six.</p>	Unidentified 40%	<p>Step 1: Identify Logical or Scientific Errors</p> <p>The question, while seemingly simple, doesn't explicitly state a constraint or limitation that might affect the racing strategy. It assumes that all horses are equally capable and that the track is always available.</p> <p>Step 2: Explain the Errors and Reasoning</p> <ul style="list-style-type: none"> Assumption of Equal Capability: The question doesn't consider that some horses might be significantly faster or slower than others. This could skew the results of a simple race. Neglect of Track Conditions: The question doesn't account for factors like weather, track condition, or horse health, which can influence performance. <p>Step 3: Reframe the Question</p>	Unidentified 40%



Q7	<p>◆ The man needs to take out all the socks except the red ones to guarantee a pair of black socks. So, the total number of socks except red ones is $21 + 1 = 22$.</p>	Unidentified 20%	<p>◆ The question contains a logical error.</p> <p>The key to solving this problem is to consider the worst-case scenario: the man could draw 21 blue socks and 31 red socks without getting a black sock. To guarantee at least one pair of black socks, he must draw one more sock.</p> <p>Therefore, the answer is 33 socks.</p>	Unidentified 40%	<p>◆ Analysis of the Question:</p> <p>1. Logical Error:</p> <p>The question seeks a guaranteed outcome, which is not feasible in this scenario. No matter how many socks the man takes out, there's always a chance that he won't draw the black sock.</p> <p>2. Explanation:</p> <ul style="list-style-type: none"> • Probability vs. Certainty: The concept of probability deals with the likelihood of an event occurring, not its absolute certainty. • Worst-Case Scenario: Even in the worst-case scenario where the man draws all the other socks first, there's still a possibility that the last sock is not black. <p>3. Reframing the Question:</p>	Unidentified 60%
Q8	<p>◆ $28000000 \text{ m/s} * 2 = 560000000 \text{ m/s}$ Luna's speed is 560000000 m/s.</p>	Unidentified 40%	<p>◆ The question contains a logical error.</p> <p>The problem lies in the statement "The spacecraft Luna is twice as fast as the Sun." This implies that Luna's speed should be $2 * 280000000 \text{ m/s} = 560000000 \text{ m/s}$.</p> <p>However, this speed exceeds the speed of light in a vacuum, which is approximately $299,792,458 \text{ m/s}$. According to the theory of relativity, no object with mass can travel faster than the speed of light.</p> <p>Therefore, the premise of the question is physically impossible and the question itself is logically flawed.</p>	Identified 100%	<p>◆ Step 1: Identify Logical or Scientific Errors</p> <p>The primary error in the question lies in the use of the word "Sun" to refer to a spacecraft. The Sun is a star, a celestial body of immense mass and energy, and it is not a spacecraft.</p> <p>Step 2: Explain the Errors and Provide Reasoning</p> <ul style="list-style-type: none"> • Misidentification of the Sun: This is a fundamental error as it confuses a celestial body with a man-made object. • Implied Constant Speed: The question doesn't specify a reference frame for the velocities. This could lead to ambiguity, as velocities are relative and can vary depending on the observer's frame of reference. <p>Step 3: Reframe the Question</p>	Unidentified 20%
Q9	<p>◆ George can remember 109,500 Chinese characters after three years.</p>	Unidentified 40%	<p>◆ The question about George learning Chinese characters has a logical flaw. It assumes that George will remember every character he learns without forgetting any. In reality, people tend to forget information over time, especially if they don't actively review it.</p> <p>Therefore, a more accurate calculation would need to account for a forgetting rate. For example, we could assume a daily forgetting rate of 5%, meaning that George forgets 5% of the characters he knows each day. This would result in a much lower total number of characters remembered after three years.</p>	Identified 60%	<p>◆ Step 1: Identify Logical and Scientific Errors</p> <p>The primary error in this question lies in its assumption of linear learning progression. It implies that George will consistently learn and remember 100 characters daily, regardless of factors like difficulty, fatigue, or the sheer volume of characters. This is not a realistic representation of human learning, especially for a complex language like Chinese.</p> <p>Step 2: Explain the Errors</p> <ul style="list-style-type: none"> • Linear Learning: Human learning is rarely linear. As the number of characters increases, the difficulty of learning and retention often rises. Factors like repetition, context, and understanding the meaning of characters influence learning efficiency. • Neglect of Forgetting Curve: The forgetting curve, a psychological principle, 	Identified 60%
Q10	<p>◆ Let's denote the side of the square as "s".</p> <p>Given, the diameter of the circle is 1.2 times the side of the square. So, the diameter = $1.2s$</p> <p>Radius of the circle: Radius = Diameter / 2 = $1.2s / 2 = 0.6s$</p> <p>Area of the circle: Area_circle = $\pi * \text{radius}^2 = \pi * (0.6s)^2 = 0.36\pi s^2$</p> <p>Area of the square: Area_square = side*2 = s^2</p> <p>Area of the square left: Area_left = Area_square - Area_circle = $s^2 - 0.36\pi s^2 = s^2(1 - 0.36\pi)$</p>	Unidentified 40%	<p>◆ The question contains a logical error.</p> <p>If the diameter of the circle is 1.2 times the side length of the square, the circle would extend beyond the boundaries of the square. This is because the maximum diameter a circle can have within a square is equal to the side length of the square.</p> <p>Therefore, the question as stated is not possible.</p>	Identified 100%	<p>◆ 1. Logical Error:</p> <p>The core logical error in the question lies in the geometric impossibility of a circle's diameter being 1.2 times the length of the square's side while fitting entirely within that square.</p> <p>2. Explanation:</p> <ul style="list-style-type: none"> • Diameter and Side Length Relationship: For a circle to fit entirely within a square, its diameter must be equal to or less than the side length of the square. This ensures that the circle touches the square's sides but doesn't extend beyond them. • 1.2 Times Relationship: A diameter 1.2 times the side length would mean the circle extends beyond the square's boundaries, making the given scenario physically impossible. <p>3. Reframed Question:</p>	Identified 100%



Appendix D



Appendix D: The result of consistent analysis

No	Response	Variation difference
Q1	 <p>To determine which toy is in the middle, let's consider the conditions:</p> <ol style="list-style-type: none"> 1. The red ball is not next to the blue truck. 2. The green dinosaur is not next to the red ball. <p>We can place the toys in a straight line and check the arrangements:</p> <ul style="list-style-type: none"> • If the red ball is at one end, it cannot be next to the blue truck, so the green dinosaur must be in the middle, and the blue truck must be at the other end. <p>Thus, the order is: blue truck - green dinosaur - red ball.</p> <p>The toy in the middle is the green dinosaur.</p>	
Q1 Variation1	<p>Thus, the arrangement is either:</p> <ul style="list-style-type: none"> • Red ball - Green dinosaur - Blue truck or • Red ball - Blue truck - Green dinosaur <p>Both satisfy the given conditions.</p>	2

Q1 Variation2		<p>Final Arrangement:</p> <ol style="list-style-type: none"> 1. Blue truck 2. Green dinosaur 3. Red ball <p>This arrangement ensures:</p> <ul style="list-style-type: none"> • The red ball is not next to the blue truck. • The red ball is not next to the green dinosaur. 		1
Q1 Variation3		<p>Final Arrangement:</p> <ol style="list-style-type: none"> 1. Blue truck 2. Green dinosaur (middle) 3. Red ball <p>Thus, the green dinosaur occupies the middle position.</p>		1

Q1 Variation4	<p>Possible Arrangements:</p> <ol style="list-style-type: none"> 1. Blue truck - Green dinosaur - Red ball 2. Green dinosaur - Blue truck - Red ball <p>These two arrangements satisfy both conditions:</p> <ul style="list-style-type: none"> • The red ball is not next to the blue truck. • The red ball is not next to the green dinosaur. 	2
Q2	 If you give away 10 apples at a time and you deliver them 13 times, you are giving away apples in batches of 10. The 12th time, like all other deliveries, would consist of 10 apples. Therefore, you gave away 10 apples the 12th time.	
Q2 Variation1	 If you give away 10 apples each time, and you deliver them 13 times, then after the 12th delivery, you would have given away: <p style="text-align: center;">$10 \text{ apples} \times 12 \text{ deliveries} = 120 \text{ apples}$</p> <p>However, since you only started with 100 apples, you wouldn't have enough apples to make 12 deliveries of 10 apples each. After 10 deliveries (giving away 100 apples), you would have no apples left. Therefore, after the 12th delivery, you would have given away all 100 apples.</p>	3

Q2 Variation2	 <p>If you had 100 apples and delivered them 13 times, giving away apples in batches of 10 each time, then:</p> <ul style="list-style-type: none"> • The total apples you would give away after 13 deliveries would be: $10 \times 13 = 130 \text{ apples}$ <p>But you only had 100 apples, which is less than the 130 apples needed for 13 deliveries. Since you can't give away more apples than you have, you would only be able to deliver apples 10 times, each time giving away 10 apples.</p> <p>Thus, you gave away 10 apples each time for the first 10 deliveries, and by the time the 10th delivery was completed, you would have no apples left. There would be no apples remaining for the 11th, 12th, or 13th delivery.</p> <p>So, the number of apples you gave away each time is 10, but you could only make 10 deliveries.</p>	3
Q2 Variation3	 <p>If you are giving away 10 apples at a time over 13 deliveries, then during the 12th delivery, you would have given away 10 apples.</p> <p>Each delivery involves giving away the same number of apples, which is 10 apples per delivery. Therefore, the amount of apples given away during the 12th delivery is 10 apples.</p>	1

Q2 Variation4	 <p>If you give away 10 apples at a time over 13 deliveries, you would have given away a total of:</p> $10 \times 13 = 130 \text{ apples}$ <p>However, you only have 100 apples. After 10 deliveries, you will have given away:</p> $10 \times 10 = 100 \text{ apples}$ <p>At this point, you have no apples left to give away. Therefore, you didn't give away any apples on the last 3 deliveries. The total apples given away on the last 3 deliveries is 0 apples.</p>	3
Q3	 <p>The sum of the interior angles of any polygon can be calculated using the formula:</p> $\text{Sum of interior angles} = (n - 2) \times 180^\circ$ <p>where n is the number of sides. For a pentagon ($n = 5$):</p> $\text{Sum of interior angles} = (5 - 2) \times 180^\circ = 3 \times 180^\circ = 540^\circ$ <p>We know the values of four of the angles: 30°, 40°, 50°, and 60°. To find the fifth angle, we subtract the sum of these four angles from 540°:</p> $\text{Fifth angle} = 540^\circ - (30^\circ + 40^\circ + 50^\circ + 60^\circ) = 540^\circ - 180^\circ = 360^\circ$ <p>So, the fifth angle is 360°.</p>	

Q3 Variation1



To find the missing fifth angle of the convex pentagon, we can use the fact that the sum of the interior angles of any convex polygon with n sides is given by the formula:

$$\text{Sum of interior angles} = 180^\circ \times (n - 2)$$

For a pentagon, $n = 5$, so the sum of the interior angles is:

$$180^\circ \times (5 - 2) = 180^\circ \times 3 = 540^\circ$$

The sum of the four known angles is:

$$30^\circ + 40^\circ + 50^\circ + 60^\circ = 180^\circ$$

To find the missing angle, subtract the sum of the known angles from the total sum of interior angles:

$$540^\circ - 180^\circ = 360^\circ$$

Thus, the missing fifth angle is 360° . However, this result is not possible for a convex pentagon because the sum of angles in a convex polygon must always be less than 360° .

2

Q3 Variation2



In any polygon, the sum of the interior angles can be calculated using the formula:

$$\text{Sum of interior angles} = (n - 2) \times 180^\circ$$

where n is the number of sides of the polygon.

For a pentagon ($n = 5$):

$$\text{Sum of interior angles} = (5 - 2) \times 180^\circ = 3 \times 180^\circ = 540^\circ$$

The sum of the given angles is:

$$30^\circ + 40^\circ + 50^\circ + 60^\circ = 180^\circ$$

To find the fifth angle, subtract the sum of the given angles from the total sum of interior angles:

$$\text{Fifth angle} = 540^\circ - 180^\circ = 360^\circ$$

Thus, the fifth angle measures **360°**.

1

Q3 Variation3



To find the size of the fifth angle in a convex pentagon, we use the formula for the sum of the interior angles of a polygon. The sum of the interior angles of a polygon with n sides is given by:

$$\text{Sum of interior angles} = (n - 2) \times 180^\circ$$

For a pentagon, $n = 5$, so the sum of the interior angles is:

$$\text{Sum of interior angles} = (5 - 2) \times 180^\circ = 3 \times 180^\circ = 540^\circ$$

Now, we know the four angles of the pentagon are 30° , 40° , 50° , and 60° . Let's add these up:

$$30^\circ + 40^\circ + 50^\circ + 60^\circ = 180^\circ$$

To find the fifth angle, subtract the sum of the given angles from the total sum of the interior angles:

$$\text{Fifth angle} = 540^\circ - 180^\circ = 360^\circ$$

So, the size of the fifth angle is **360°** .

1

Q3 Variation4



To find the largest angle in the convex pentagon, we can use the property that the sum of the interior angles of any polygon with n sides is given by the formula:

$$\text{Sum of interior angles} = (n - 2) \times 180^\circ$$

For a pentagon ($n = 5$):

$$\text{Sum of interior angles} = (5 - 2) \times 180^\circ = 3 \times 180^\circ = 540^\circ$$

We are given the four angles of the pentagon: 30° , 40° , 50° , and 60° . To find the missing angle, we subtract the sum of these four angles from the total sum of the interior angles:

$$\text{Missing angle} = 540^\circ - (30^\circ + 40^\circ + 50^\circ + 60^\circ)$$

First, sum the given angles:

$$30^\circ + 40^\circ + 50^\circ + 60^\circ = 180^\circ$$

Now, subtract:

$$\text{Missing angle} = 540^\circ - 180^\circ = 360^\circ$$

Thus, the largest angle in the pentagon is 360° .

1

Q4



To find how many patients did not recover, we need to subtract the number of patients who recovered fully and the number of patients who died from the total number of patients treated.

Here's the calculation:

- Total patients treated = 98.4
- Fully recovered patients = 43.7
- Patients who died = 10

So, the number of patients who did not recover (including those who died) is:

$$98.4 - 43.7 - 10 = 44.7$$

Thus, 44.7 patients did not recover.

Q4 Variation1



To find out how many patients did not recover, we need to subtract the number of patients who made a full recovery and the number of patients who died from the total number of patients treated.

We know:

- Total patients treated = 98.4
- Patients who made a full recovery = 43.7
- Patients who died = 10

The number of patients who did not recover is:

$$\text{Patients who did not recover} = 98.4 - 43.7 - 10 = 44.7$$

So, 44.7 patients did not recover.

1

Q4 Variation2



To find how many patients did not recover, you can subtract the number of patients who fully recovered and the number of patients who died from the total number of patients treated.

Given:

- Total patients treated = 98.4
- Fully recovered = 43.7
- Patients who died = 10

Patients who did not recover = Total patients treated - (Fully recovered + Patients who died)


$$= 98.4 - (43.7 + 10)$$


$$= 98.4 - 53.7$$

$$= 44.7$$

So, **44.7 patients** did not recover.

1

Q4 Variation3	<p>Therefore:</p> <ul style="list-style-type: none"> • Recovered patients: 43.7 • Patients who did not recover: 54.7 <p>To compare these two groups, we can calculate the ratio:</p> $\text{Ratio of recovered to not recovered} = \frac{43.7}{54.7} \approx 0.798$ <p>This means that for every patient who recovered, there were approximately 0.8 patients who did not recover.</p>	1
Q4 Variation4	<p> To find how many more patients recovered than died, subtract the number of deaths from the number of recoveries:</p> $43.7 \text{ (recovered patients)} - 10 \text{ (deaths)} = 33.7$ <p>So, 33.7 more patients recovered than died.</p>	1

Q5	<div data-bbox="696 250 1585 817">  <p>Let's break down the information step by step:</p> <ol style="list-style-type: none"> A finished before B, but behind C. This means the order is: C, A, B. D finished before E, but behind B. This means the order so far is: C, A, B, D, E. E finished before A, but behind D. This means the order must be: C, D, E, A, B. <p>So, the finishing order is: C, D, E, A, B.</p> </div>
----	---

Q5 Variation1	<p>We can deduce the following order by fitting them together:</p> <ol style="list-style-type: none"> 1. C finished first (since C is before A and B in the first condition). 2. D finished second (since D is after B but before E, and also before A in the third condition). 3. E finished third (since E is after D but before A). 4. A finished fourth (since A is after C but before B, and after E). 5. B finished fifth (since B is last in the first condition and before D in the second condition). <p>So, the finishing sequence is:</p> <p>C, D, E, A, B.</p>	1
---------------	---	---

Q5 Variation2



Let's analyze the information step by step:


- **A finished after C but before B:** This means C is before A, and A is before B. C_A_B
- **D finished after B but before E:** This means B is before D, and D is before E. $C_A_B_D_E$
- **E finished after D but before A:** This means D is before E, and E is before A. $C_D_E_A_B$

Now we can deduce the order from the constraints:

1. C finished first.
2. D finished second.
3. E finished third.
4. A finished fourth.
5. B finished fifth.

Therefore, **A finished in 4th position.**

1

Q5 Variation3	 <p>To solve this puzzle, we need to analyze the positions of the participants based on the given conditions:</p> <ol style="list-style-type: none"> 1. A finished between C and B: This implies A is somewhere between C and B in the ranking. So, the order must be either C, A, B or B, A, C. 2. D finished between B and E: D is somewhere between B and E. So, the order must be either B, D, E or E, D, B. 3. E finished between D and A: E is somewhere between D and A. So, the order must be either D, E, A or A, E, D. <p>Now, let's combine these pieces of information.</p> <ul style="list-style-type: none"> • From condition 1, we know A is between C and B. • From condition 2, D is between B and E. • From condition 3, E is between D and A. <p>The order that satisfies all conditions is: C, A, B, D, E.</p> <p>Thus, E is the last one to finish.</p>	3
---------------	--	---

Q5 Variation4	<p>Let's analyze how these clues fit together:</p> <ul style="list-style-type: none"> • Start with clue 2 (D between B and E): the possible orders are B, D, E or E, D, B. • If we take the first option, B, D, E, then clue 3 tells us that E is between D and A, so we get D, E, A. • Now, from clue 1 (A between C and B), we know the order for C, A, B is either C, A, B or B, A, C. • Combining all clues, the only consistent sequence is: C, A, B, D, E. <p>So, the final order in which the individuals finished is: C, A, B, D, E.</p>	3
---------------	--	---