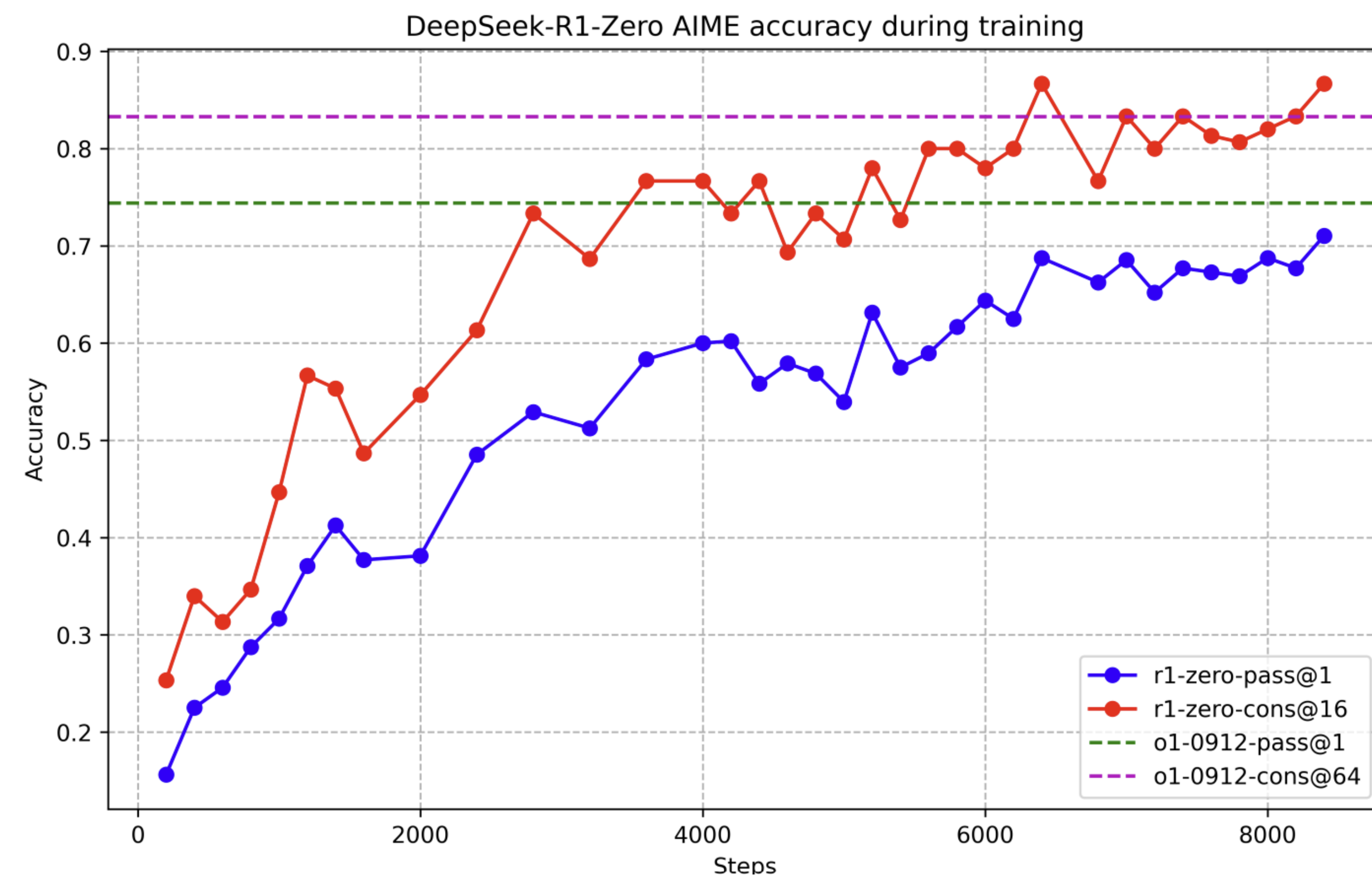
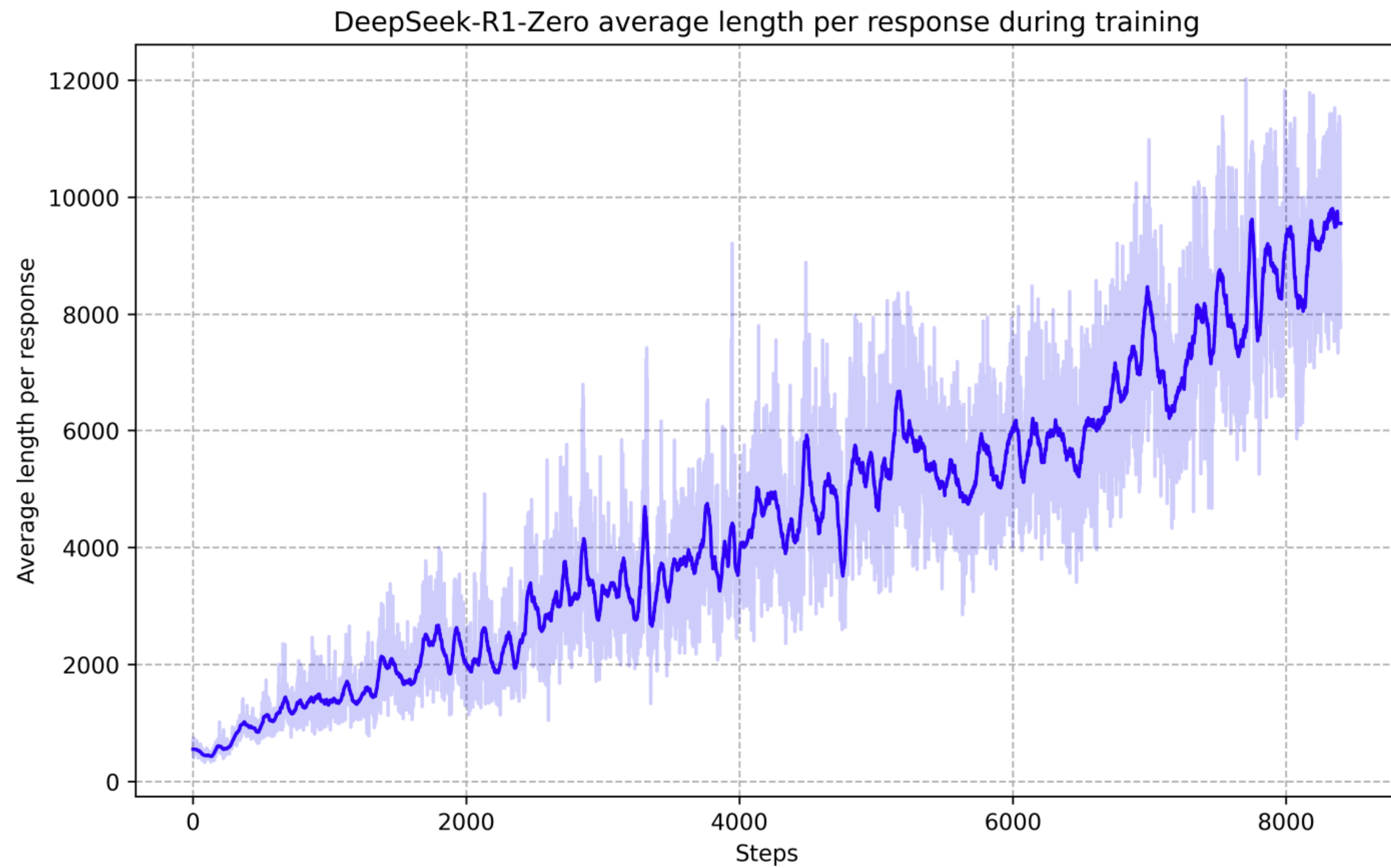


More Thought, Less Accuracy? On the Dual Nature of Reasoning in Vision-Language Models

Xinyu Tian

Australian National University

Motivation



Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a+x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a+x}} = x$, let's start by squaring both \dots

$$\left(\sqrt{a - \sqrt{a+x}}\right)^2 = x^2 \implies a - \sqrt{a+x} = x^2.$$

Rearrange to isolate the inner square root term:

$$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$$

\dots

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be \dots

We started with the equation:

$$\sqrt{a - \sqrt{a+x}} = x$$

First, let's square both sides:

$$a - \sqrt{a+x} = x^2 \implies \sqrt{a+x} = a - x^2$$

Next, I could square both sides again, treating the equation: \dots

\dots

Successful Practices in LMs:

- **RL brings prolonged and advanced reasoning capability.**
- **The improved performance is driven by the evolved reasoning.**

Motivation

How many r in raspberry? **Question**

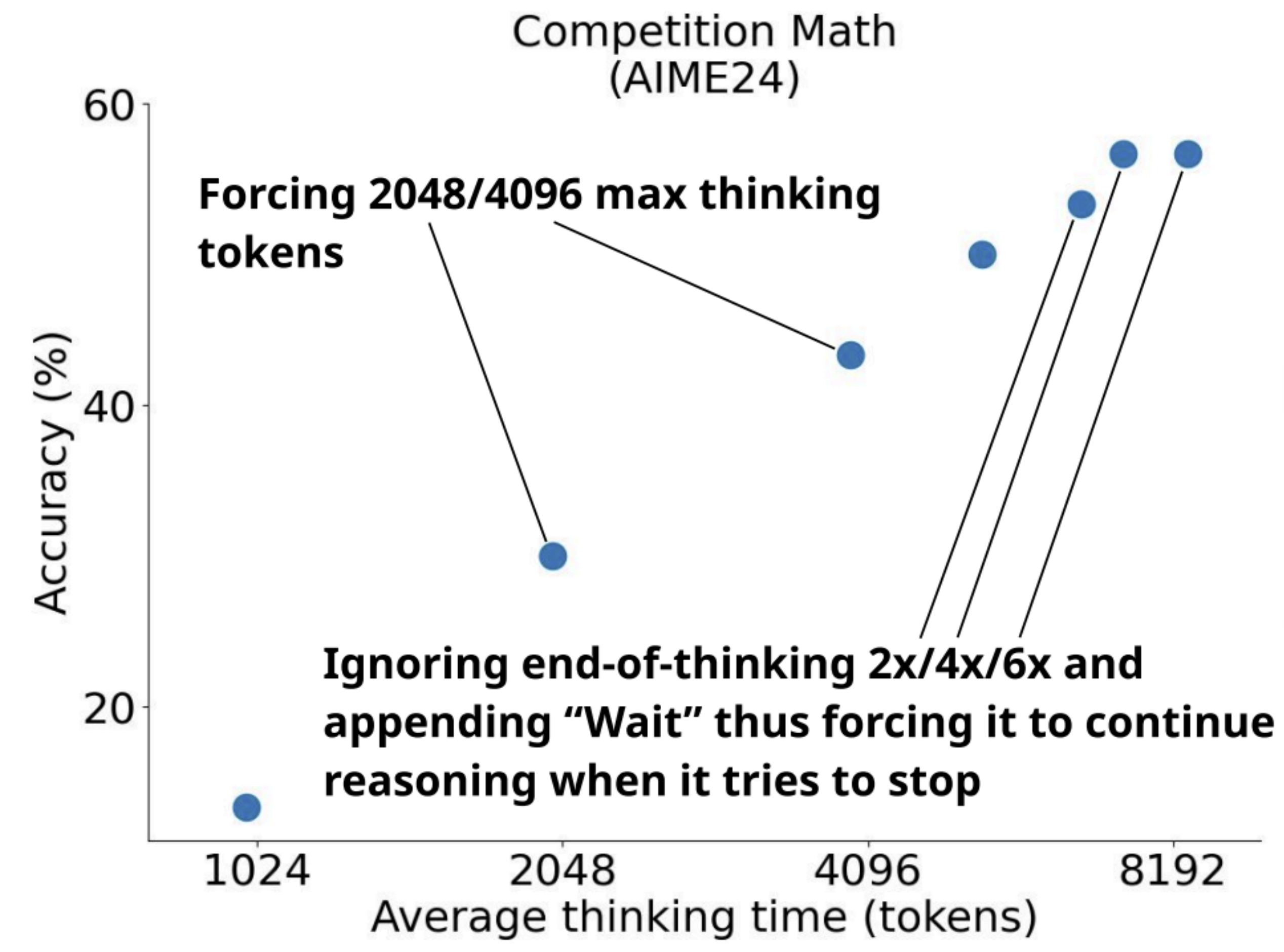
Let's break down the process of counting the letter 'r' in the word "raspberry" ...

- * First letter: 'r' - This is an 'r', count = 1.
- * Second letter: 'a' - Not an 'r', count remains 1 ...
- * Sixth letter: 'e' - Not an 'r', count remains 1.
- * Seventh letter: 'r' - This is an 'r', count = 2.
- * Eighth letter: 'y' - Not an 'r', count remains 2 ...

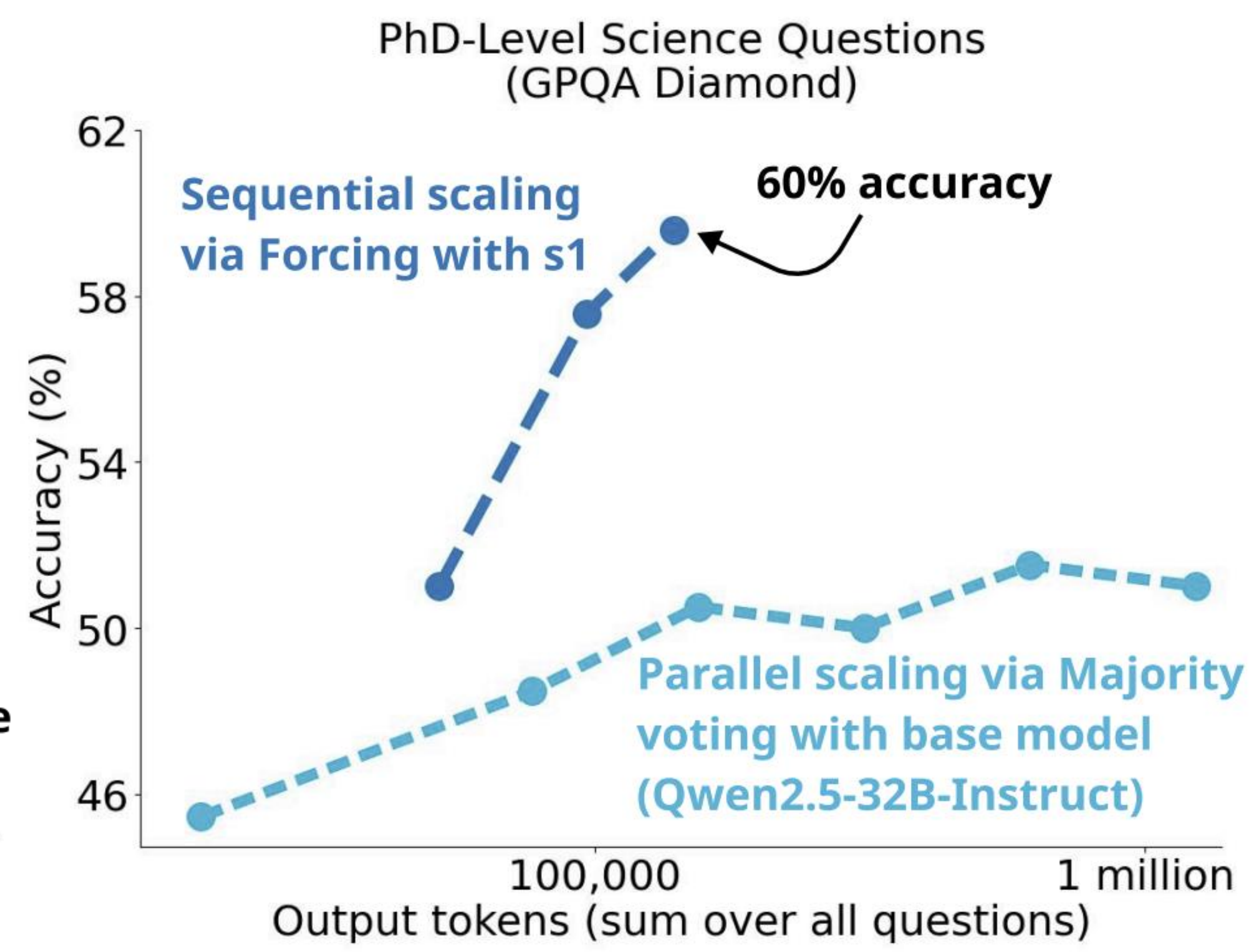
The number of 'r's in "raspberry" is 2.

Wait, let's re-read the question carefully. It asks "How many r in raspberry?" ... * r - a - s - p - b - e - r - r - y ... * First 'r' ... * Second 'r' ... * Third 'r' ... Count = 3 ... **Reasoning trace**

My initial answer of 2 was incorrect due to a quick reading of the word. **Final Answer:** The final answer is **3** **Response**



(a) Sequential scaling via budget forcing



(b) Parallel scaling via majority voting

- **SFT on distilled reasoning traces also brings competitive performance for LMs**
- **Test-time scaling laws: more reasoning/thought will bring higher benefits**

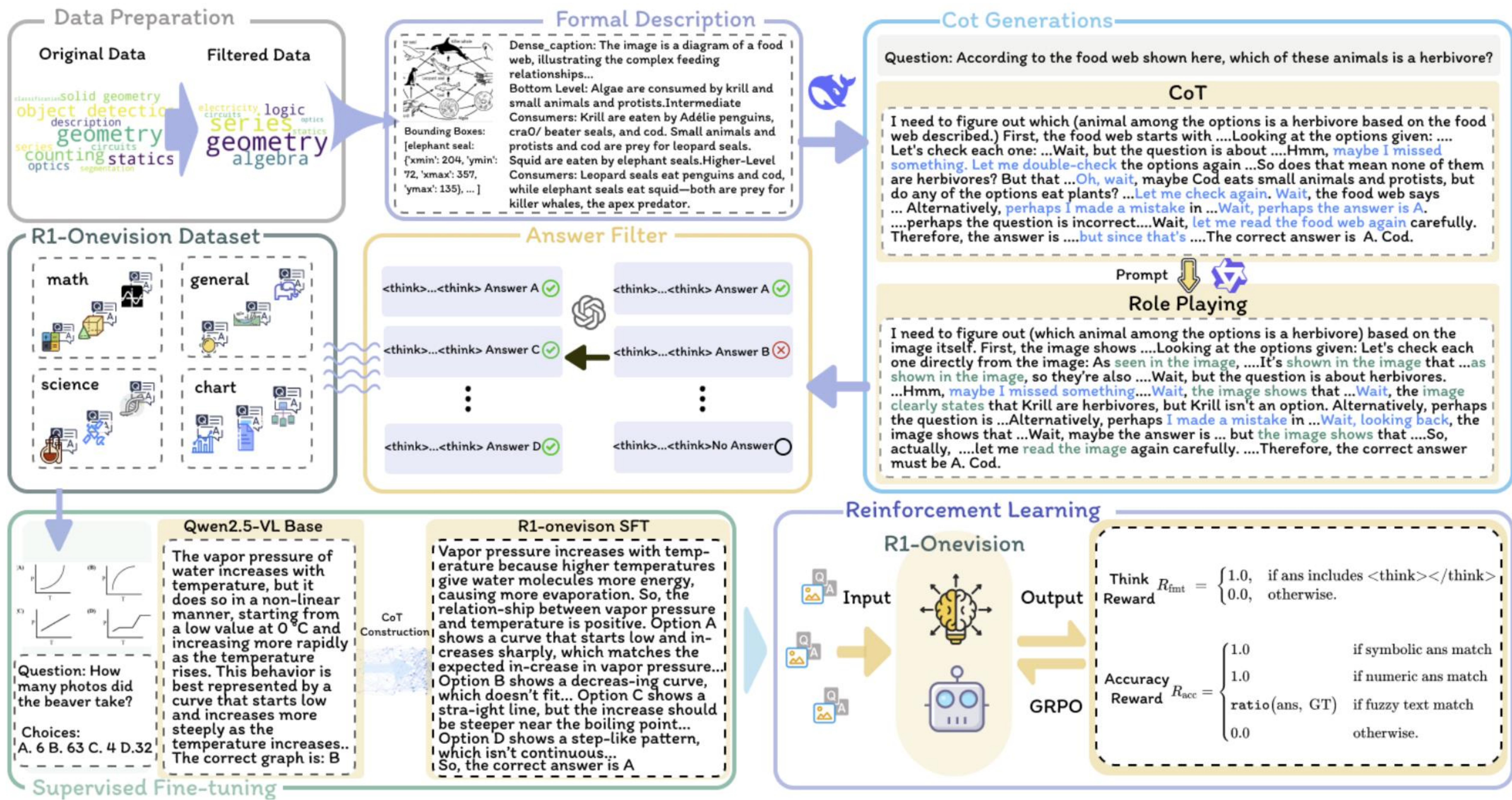
Motivation

Question: Can we bring such strong reasoning capability into multi-modal domains?

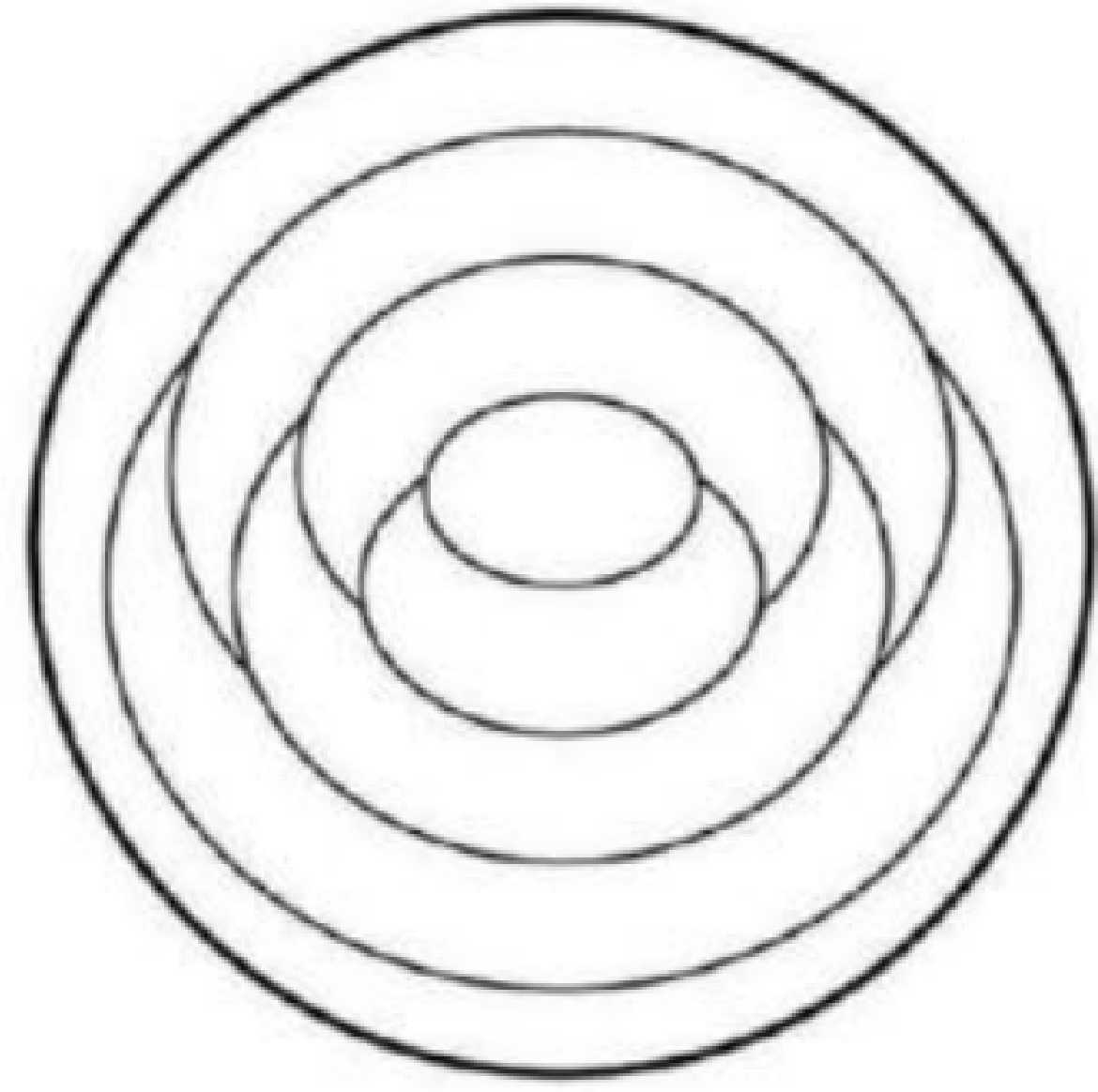
A lot of recent works are pursuing strong multi-modal reasoning models

- **R1-OneVision** *Yang, Yi, et al. "R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization." arXiv preprint arXiv:2503.10615 (2025).*
- **Visual-RFT** *Liu, Ziyu, et al. "Visual-rft: Visual reinforcement fine-tuning." arXiv preprint arXiv:2503.01785 (2025).*
- **Reason-RFT** *Tan, Huajie, et al. "Reason-rft: Reinforcement fine-tuning for visual reasoning." arXiv preprint arXiv:2503.20752 (2025).*
- **VLAA-Thinker** *Chen, Hardy, et al. "Sft or rl? an early investigation into training r1-like reasoning large vision-language models." arXiv preprint arXiv:2504.11468 (2025).*
- **Vision-R1** *Huang, Wenxuan, et al. "Vision-r1: Incentivizing reasoning capability in multimodal large language models." arXiv preprint arXiv:2503.06749 (2025).*
- **ThinkLite-VL** *Wang, Xiyao, et al. "Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement." arXiv preprint arXiv:2504.07934 (2025).*
- **SEED1.5-VL** *Guo, Dong, et al. "Seed1. 5-vl technical report." arXiv preprint arXiv:2505.07062 (2025).*
- **Gemini-2.5-Pro** *Comanici, et al. "Gemini 2.5: Pushing the frontier with advanced reasoning, and next generation agentic capabilities." arXiv preprint arXiv:2507.06261 (2025).*
- **GPT-5**
- ...

Motivation



Motivation



Question: Cynthia paints each region of the figure in a single color: red, blue or yellow. She paints with different colors the regions that touch each other. In how many different ways can Cynthia paint the figure?



GPT-4o Caption:

The image features a series of concentric circles with varying thicknesses, forming a symmetrical pattern. The outermost circle is the largest, enclosing progressively smaller circles toward the center. Some of the inner circles intersect with each other, creating curved segments that resemble a layered or nested design. The overall composition appears structured and balanced, with smooth, evenly spaced curves contributing to a visually appealing geometric arrangement. The image is monochromatic, utilizing black outlines on a white background.



Deepseek-R1:

Okay, let's tackle this problem step by step...First, I need to visualize the figure described...Since it's concentric circles, ...The challenge is to figure out how many regions there are and their adjacency. Let me try to break it down:...**Let's consider possible common figures of this type...**Case 1: A simple target with three concentric circles...**Total: $3*2*2*2 = 24$.** Case 2: Suppose there are two intersecting circles inside...**Total: $3*2*2=12$.** However, I recall that in similar problems with a central region and surrounding regions forming a cycle...Assuming the structure forms a chain of outer, middle, and inner annuli, plus four inner regions arranged in a cycle the coloring constraints are:

1. Outer annulus (O): 3 color choices.
2. Middle annulus (M): **2 choices (different from O).**
3. Inner annulus (I): **2 choices (different from M).**
4. Four inner regions: **Form a cycle, requiring 2 alternating colors different from I, yielding 2 colorings.**

Total colorings:

$3(O)*2(M)*2(I)*2(\text{inner cycle})=24$. Final answer: 24.



Qwen2.5-VL-7B:

To determine the number of different ways Cynthia can paint the figure, we need to consider the constraints given: each region must be painted with a different color from its adjacent regions. Let's analyze the figure step by step.

1. Identify the regions.
2. Coloring the innermost circle
3. Coloring the next circle: The next circle must be a different color from the innermost circle. **So, there are 2 choices for this circle.**
4. Coloring the next circle: The third circle must be a different color from the second circle. So, there are 2 choices for this circle.
5. Coloring the next circle
6. Coloring the outermost circle - 3 choices for the innermost circle.

- 2 choices for the second circle. 2 choices for the third circle. 2 choices for the fourth circle. 2 choices for the outermost circle.

The total number of different ways to paint the figure is:

$3*2*2*2*2 = 3*16 = 48$. Thus, the number of different ways Cynthia can paint the figure is **48**.



R1-Onevision:

The image shows a figure with five concentric circles, each touching the next. Cynthia uses three colors: red, blue, and yellow. Each circle must be painted a different color from its neighbors. Since the circles are concentric, each circle touches the next, meaning each circle must be painted a different color.

Let's analyze the coloring:

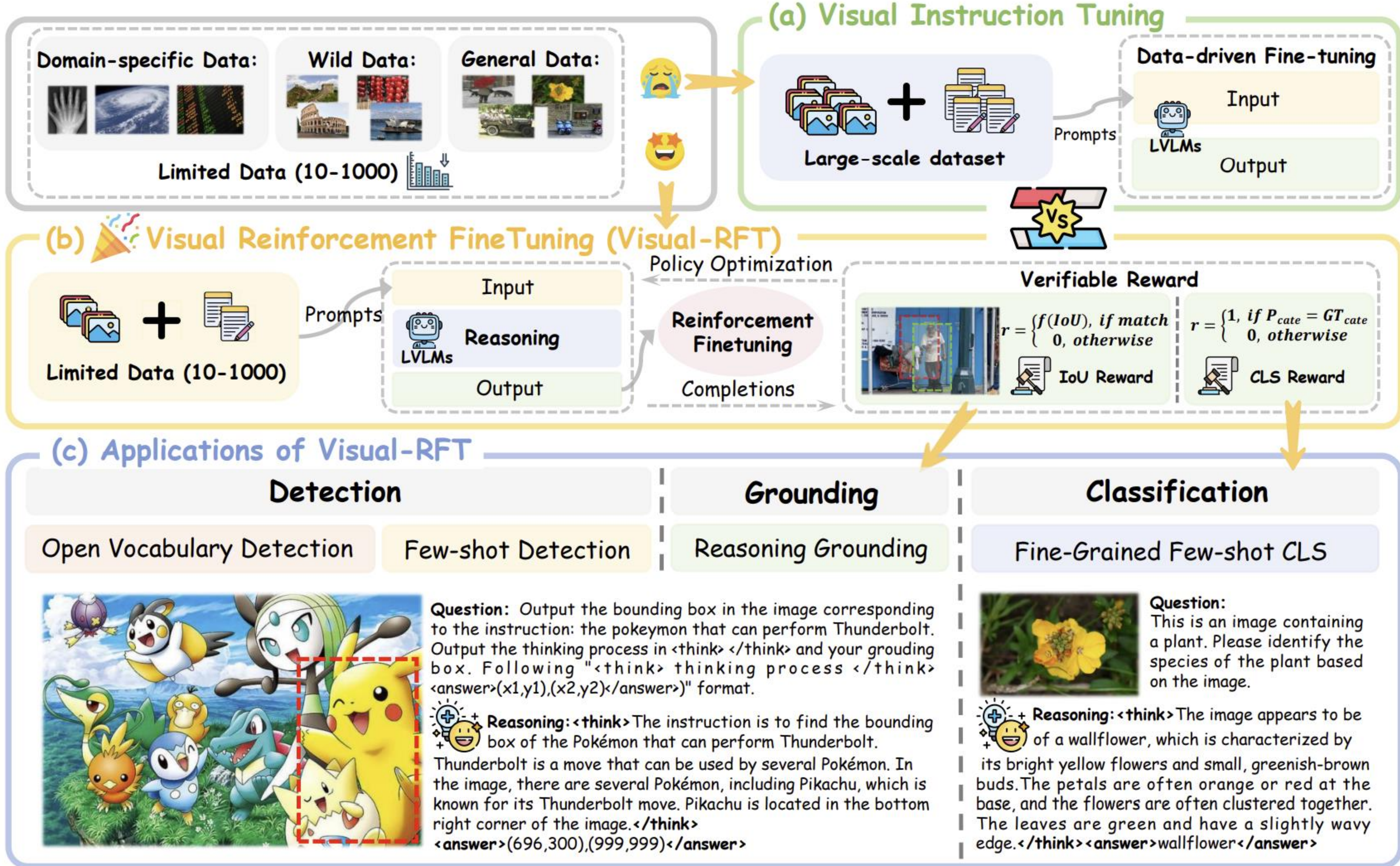
1. The innermost circle can be any of the three colors (red, blue, or yellow).
2. The next circle must be a different color from the innermost circle.
3. The third circle must be a different color from the second circle.
4. The fourth circle must be a different color from the third circle.
5. The outermost circle must be a different color from the fourth circle.

Given that there are three colors and each circle must be a different color from its neighbors, the number of ways to color the circles is the number of permutations of the three colors. **This is calculated as $3!$ (3 factorial), which is 6. Therefore, the number of different ways Cynthia can paint the figure is 6.**

The number of different ways Cynthia can paint the figure is **6**.

- More advanced reasoning behaviors, e.g., self-reflection, verification
- Less reasoning errors, e.g., calculation, consistency

Motivation



Motivation



This is an image containing a flower. Please identify the species of the flower based on the image



<answer> **globe-flower** </answer>



<think> The image shows a yellow buttercup flower with a green stem and green leaves in the background. Buttercups are a type of flowering plant in the buttercup family. They are known for their bright yellow flowers and are commonly found in meadows and grasslands. </think>
<answer> **buttercup** </answer>



This is an image containing a pet. Please identify the species of the pet based on the image



<answer> **hedgehog** </answer>



<think> The dog in the image appears to be a Japanese Chin, also known as a Japanese Spaniel. This breed is known for its small size, fluffy coat, and distinctive black and white coloration. The Japanese Chin is a popular pet and is known for its friendly and affectionate nature. </think>
<answer> **Japanese Chin (or Japanese Spaniel)** </answer>



- Reasoning can also improve general visual tasks, e.g., classification

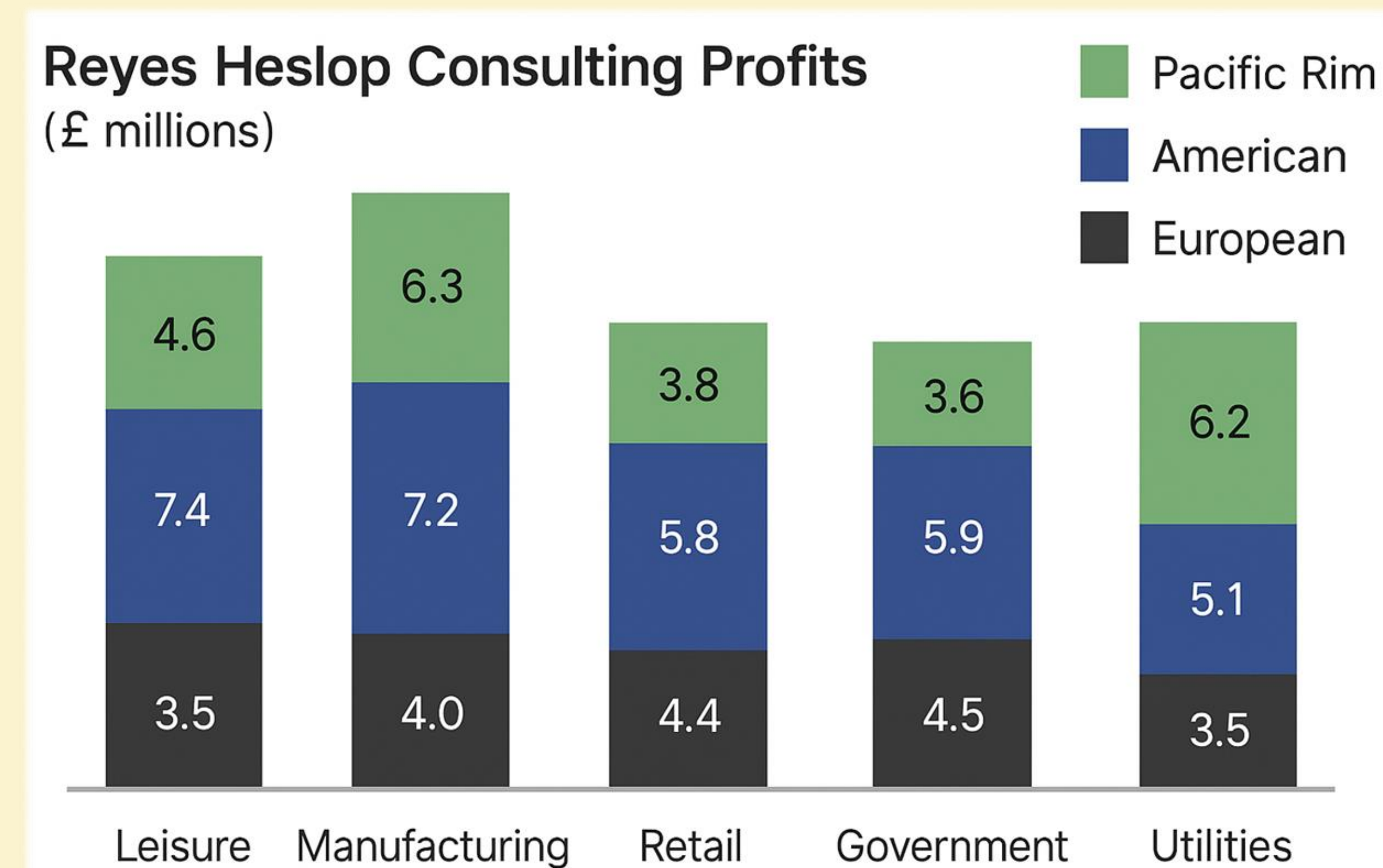
Motivation

It seems that reasoning is a universal remedy to various visual tasks.

However, does more reasoning always help VLMs accomplish their tasks?

LogicVista (Chart)

Q: What is the total profit of top three industries in America? GT Answer: 20.5



Full Reasoning

<think> We are given a chart of the profit from Reyes Heslop, and our aim is to calculate total profit of top three industries in America. Let's identify the profits of each industry:

1. Leisure: 4.6 (Pacific Rim), 7.4 (American) and 3.5 (European)
2. Manufacturing: 6.3 (Pacific Rim), 7.2 (American) and 4.0 (European)
3. ...

4. Utilities: 6.2 (Pacific Rim), 8.1 (American) and 3.5 (European)

Back to the question, the top three industries are Leisure (7.4), Manufacturing (7.2), and Utilities (8.1). Let's sum the profit up: $7.4 + 7.2 + 8.1 = 22.7$ (millions).</think>

<answer>The total profit is 22.7 millions.</answer>

Early Decision

<think> We are given a chart of the profit from Reyes Heslop, and our aim is to calculate total profit of top three industries in America. Let's identify the profits of each industry:

1. Leisure: 4.6 (Pacific Rim), 7.4 (American) and 3.5 (European)

Time is up. I should stop thinking and formulate a final answer based on what I have.</think>

<answer>The sum of top three industries in America is 20.5 millions</answer>

- We control the reasoning length by letting VLMs decide their answers in advance
- This helps us monitor the contribution of different reasoning parts to performance

Motivation

It seems that reasoning is a universal remedy to various visual tasks.

However, does more reasoning always help VLMs accomplish their tasks?

MMStar (Counting)

Q: In this given image, how many white cats are there can you see in total? **GT Answer:** 4



Full Reasoning

<think> Okay, we are asked how many cats are there in the photo. Let's have a look.

1. First, there is an adult white cat lying down with distinct blue and yellow eyes.
2. On the far left, a kitten lies on its back with tiny paws stretched upwards.
3. There is one cat in the middle-bottom of the photo, with head tilted to the side.
4. To the right, closest to the mother's head, a kitten is resting with belly showing.
5. On the far right, there is a kitten hidden beneath, with only its black eyes visible.

<answer>The number of cats is 5</answer>

Early Decision

<think> Okay, we are asked how many cats are there in the photo. Let's have a look.

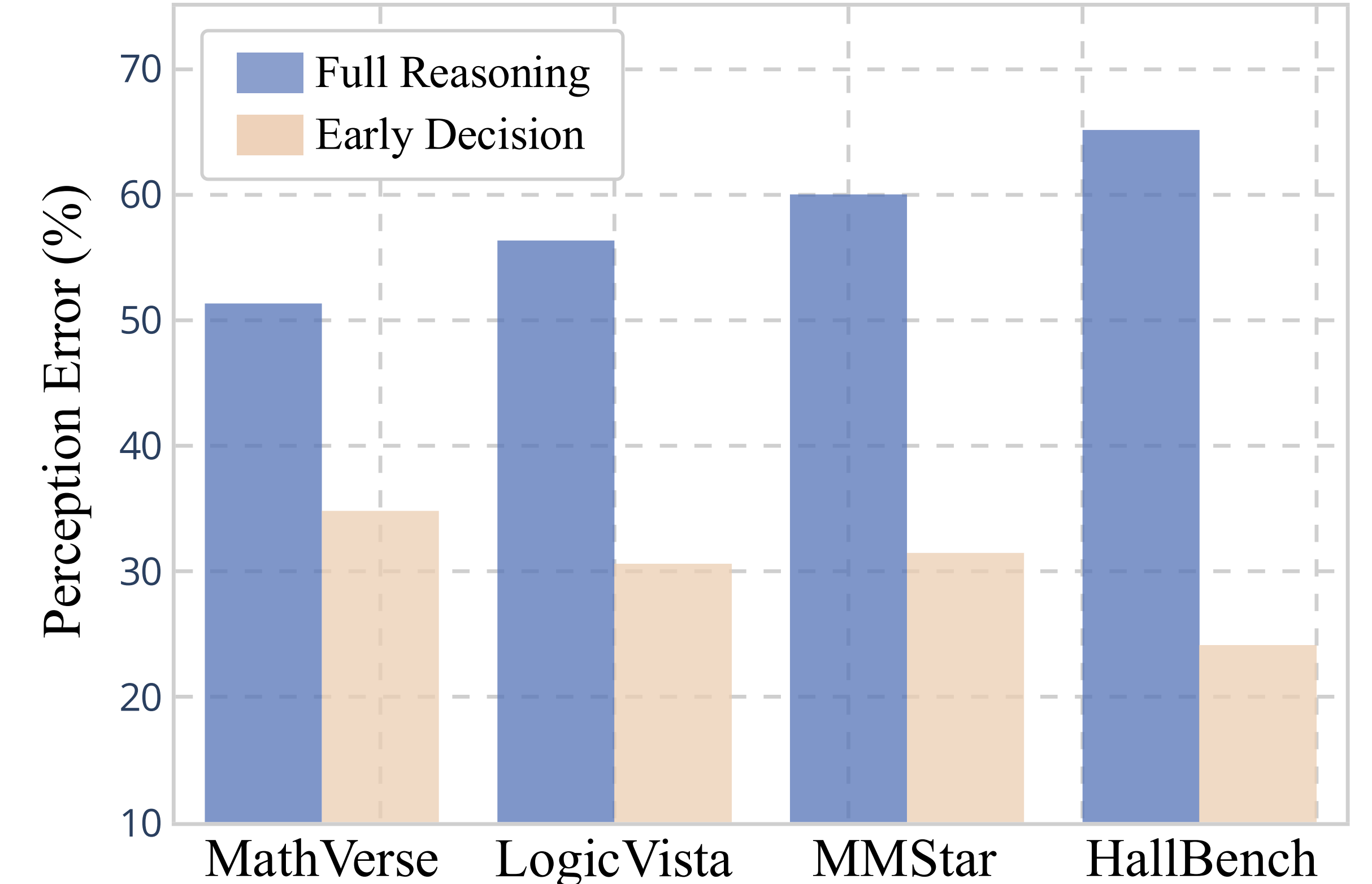
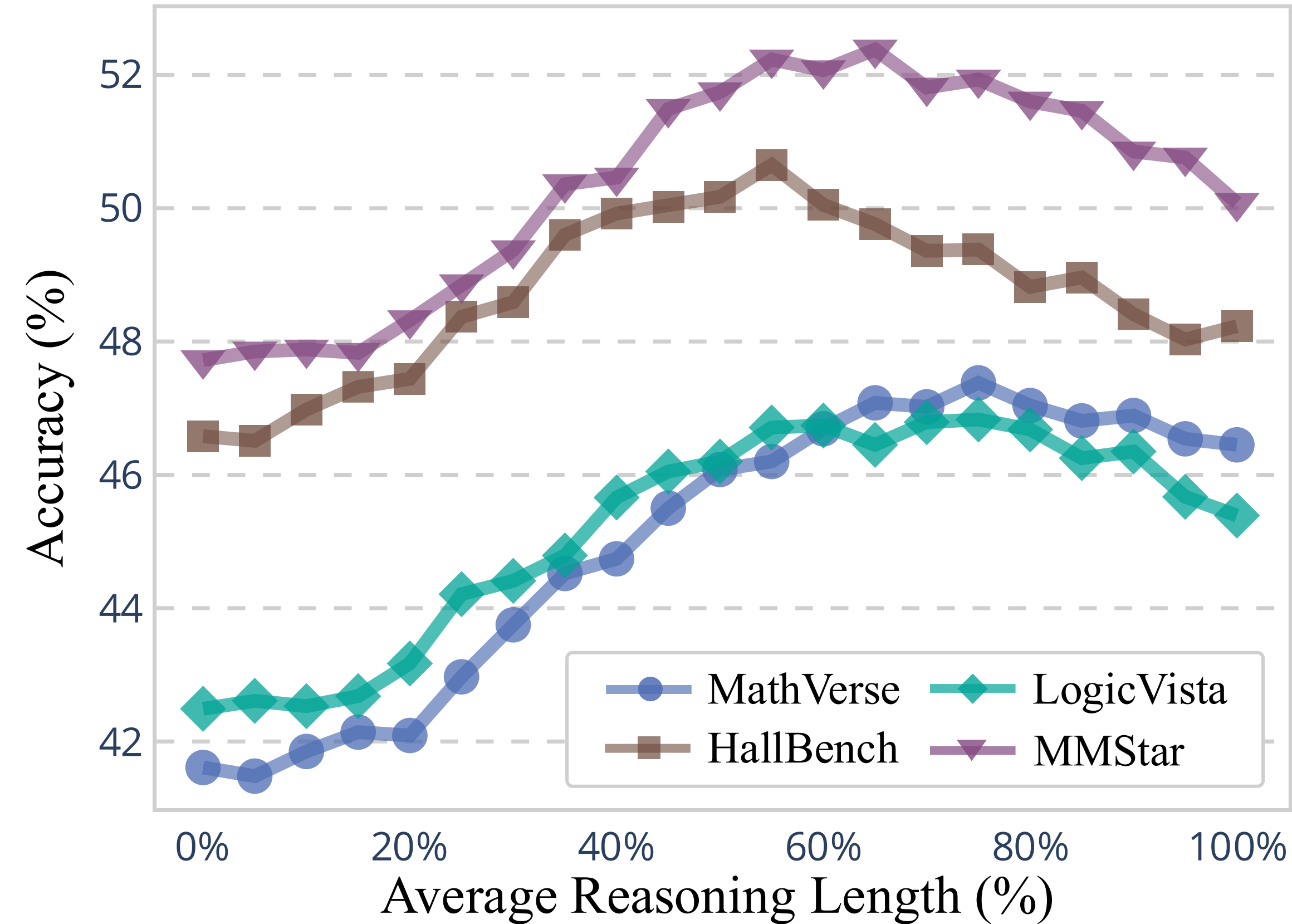
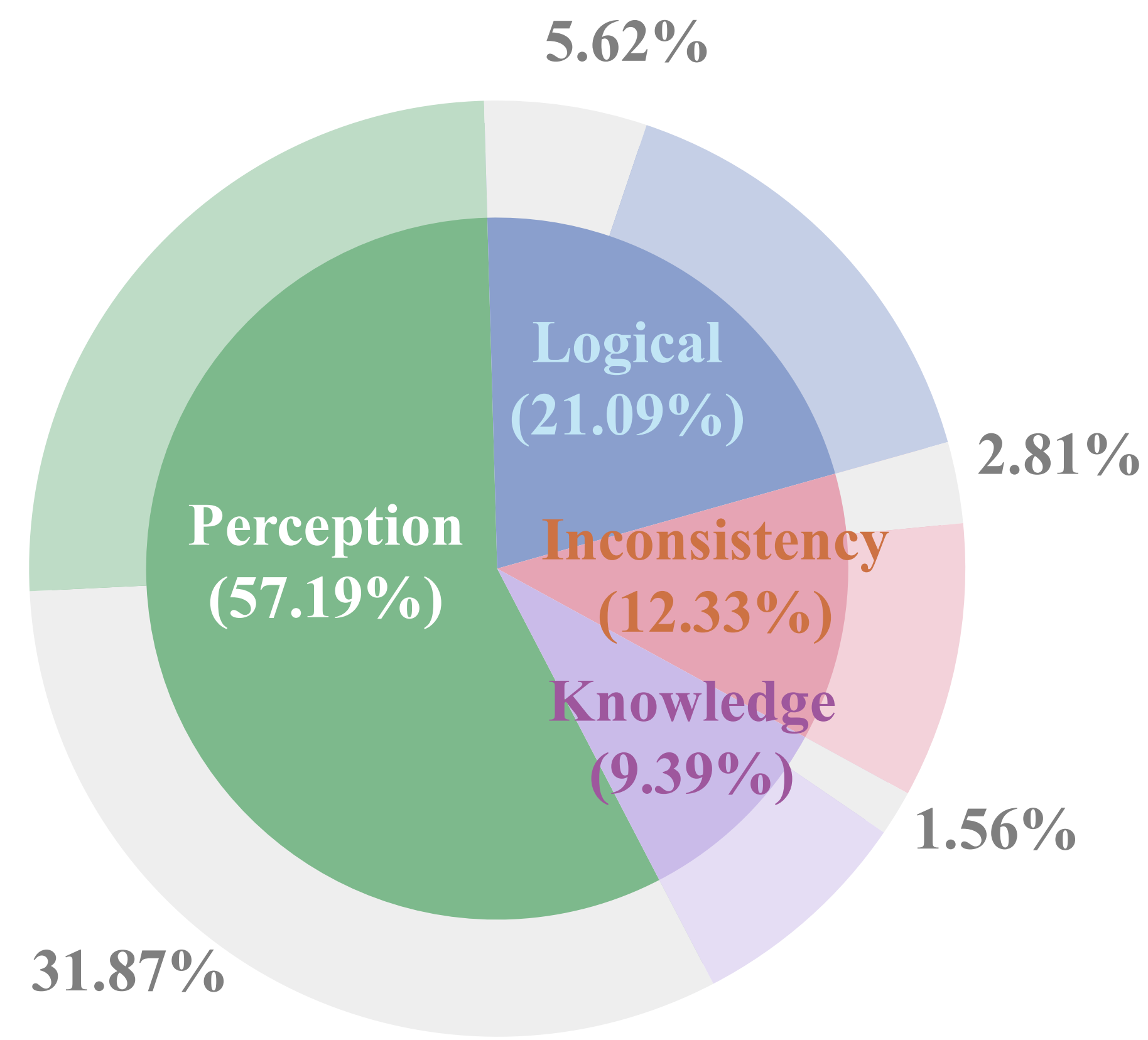
1. First, there is an adult white cat lying down with distinct blue and yellow eyes.
2. On the far left, a kitten lies on its back with tiny paws stretched upwards.

Time is up. I should stop thinking and formulate a final answer based on what I have.</think>

<answer>There are 4 cats in this photo</answer>

- For many problems where VLMs make mistakes under full reasoning, an early decision can instead let models reveal the correct answers.
- This leads us to question whether reasoning is truly a constant boost to accuracy.

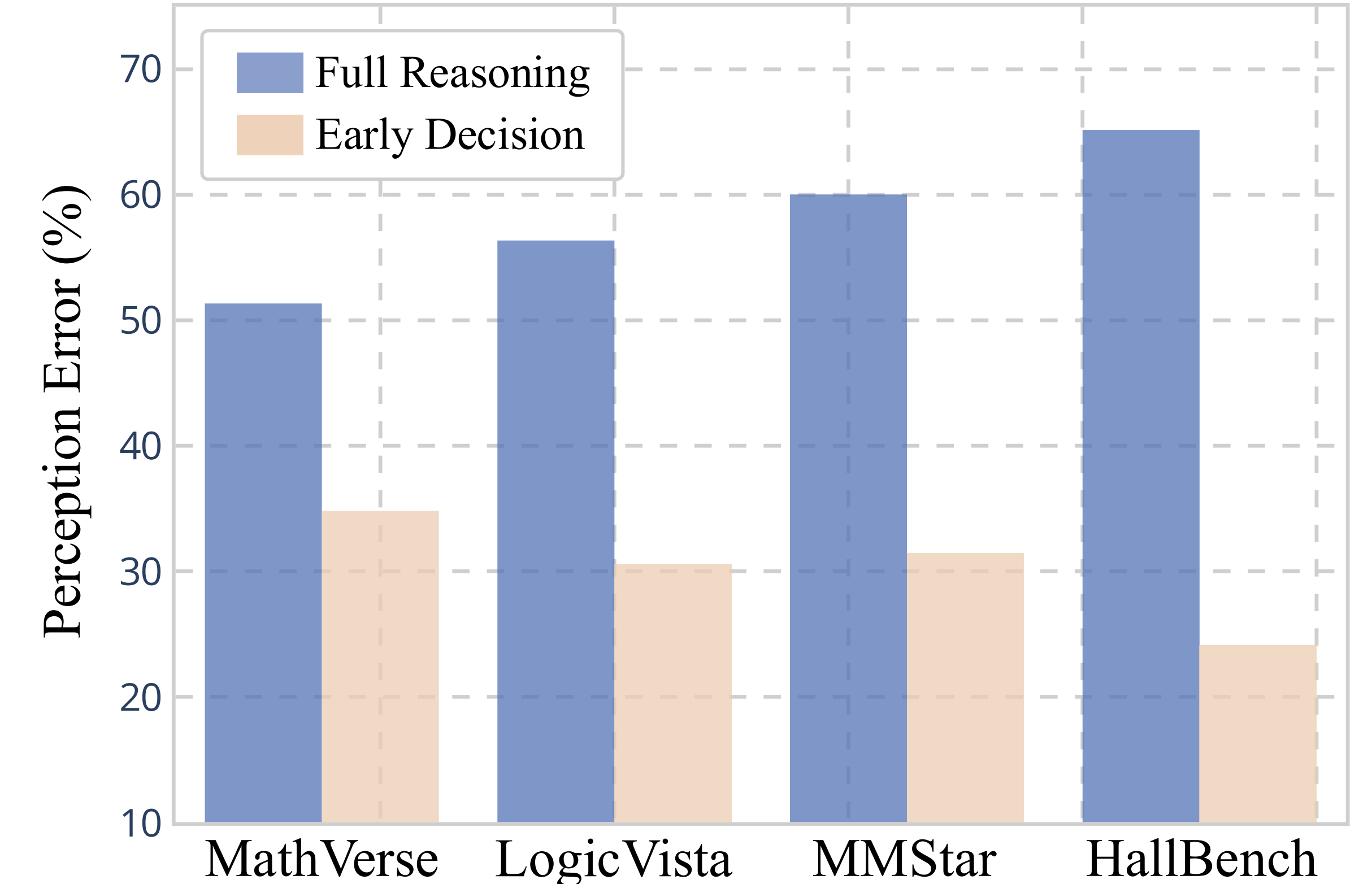
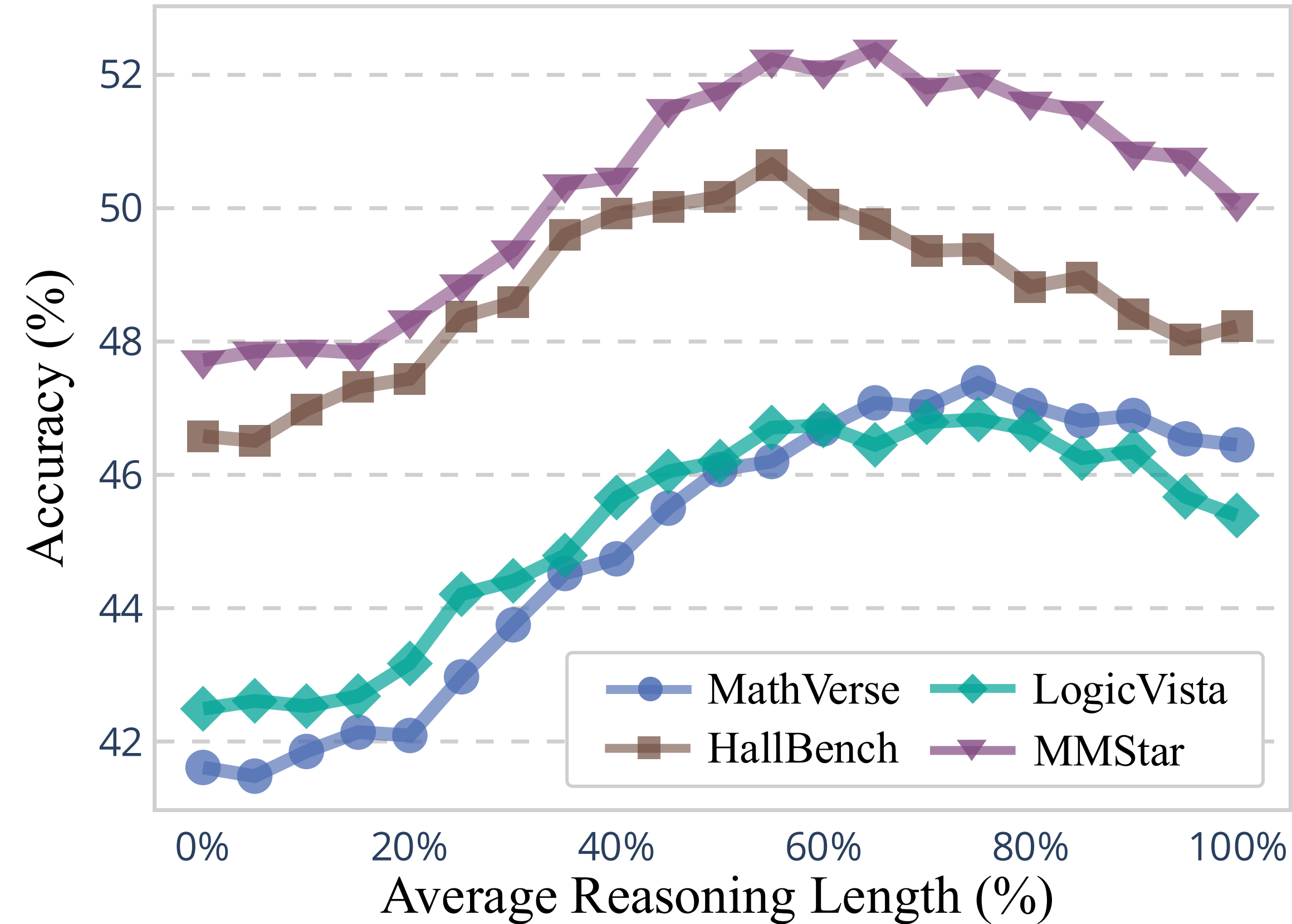
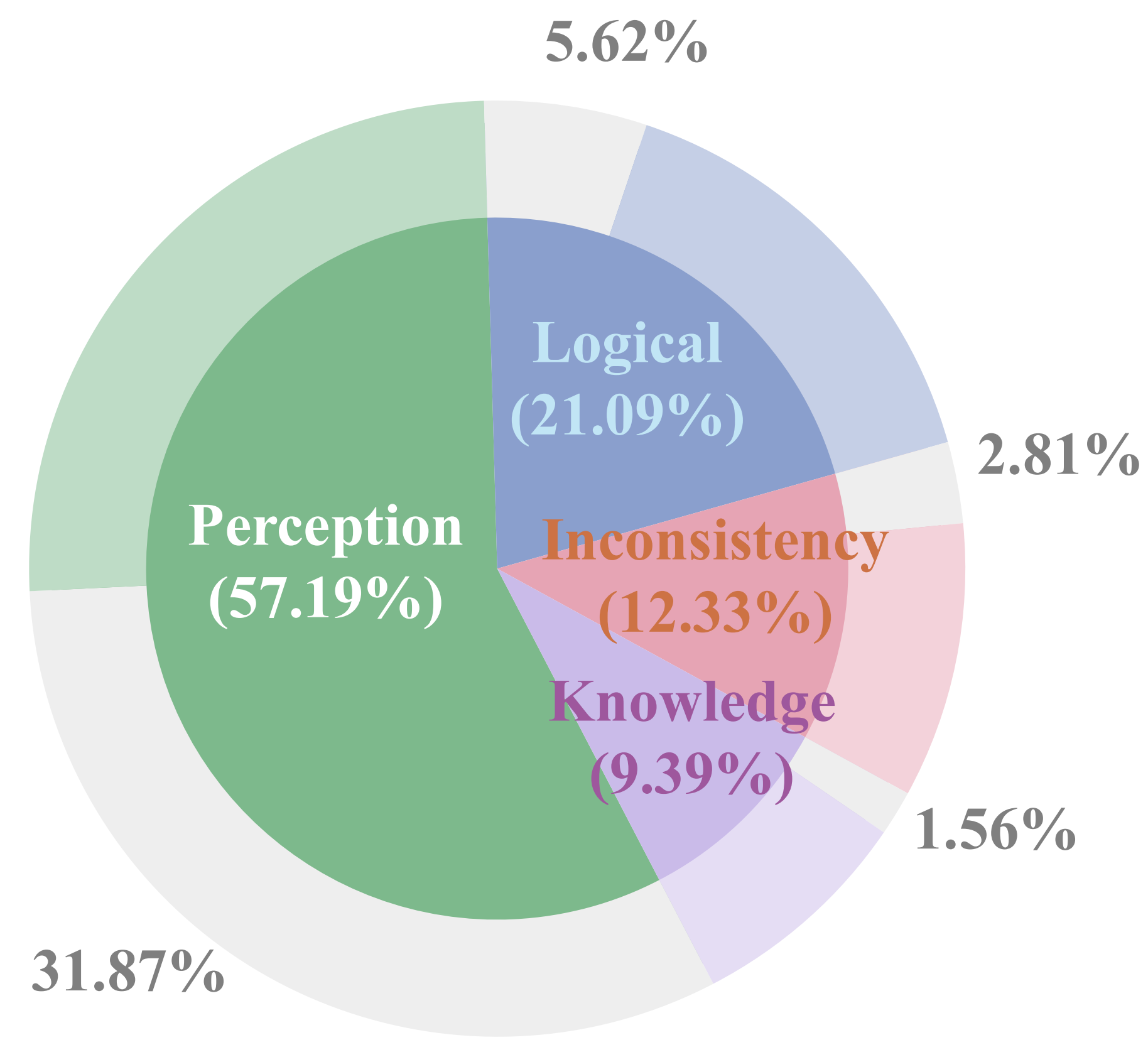
Dual Nature of Reasoning



- We first use GPT-4 to determine multiple appropriate positions to terminate reasoning and let models give their answers in advance (early stopping), and plot accuracy over varying reasoning length on multiple benchmarks.
- We find that longer reasoning is not always better; instead, there is a peak, after which the model performance begins to slightly decline with the reasoning progressing.

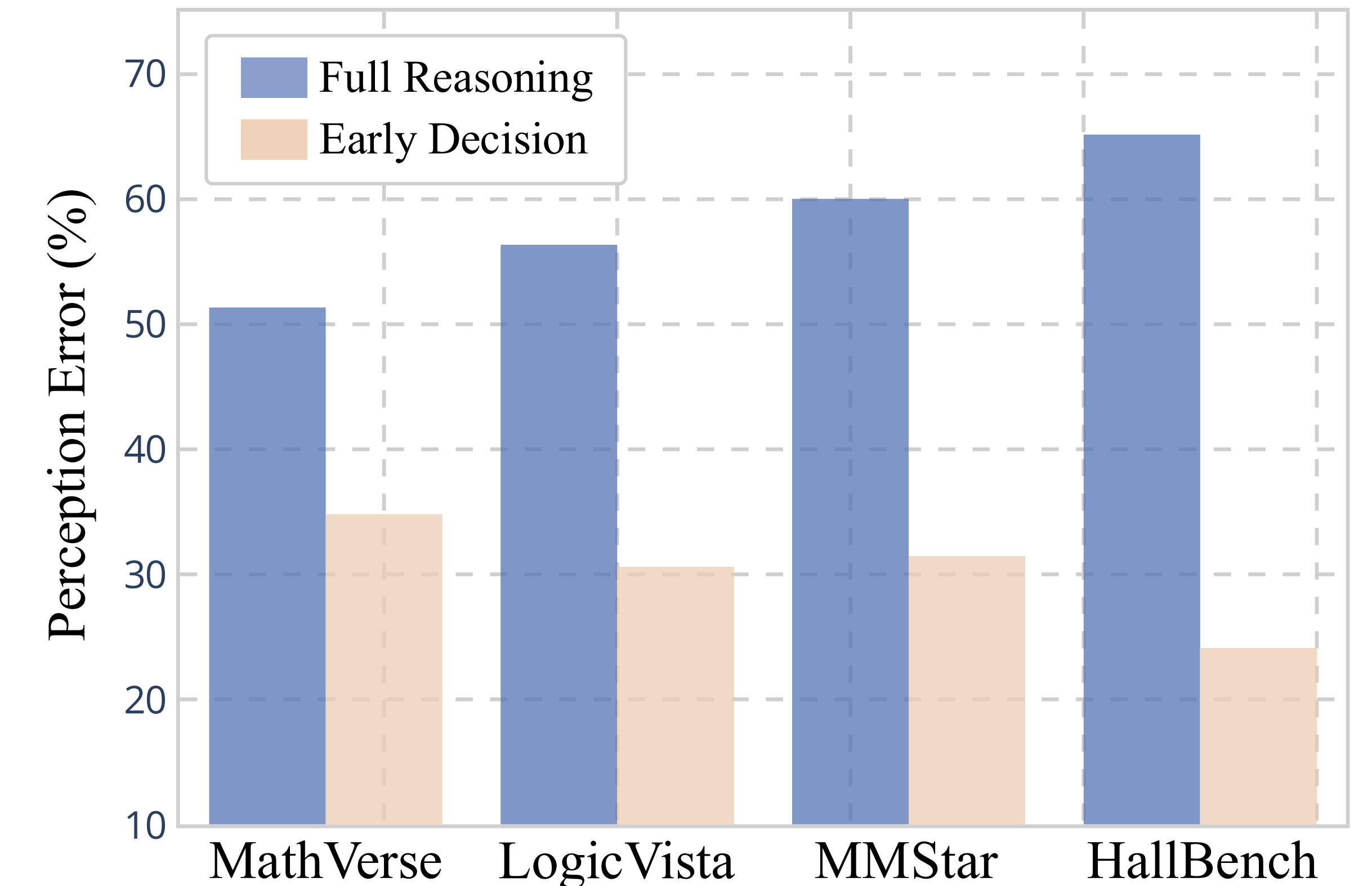
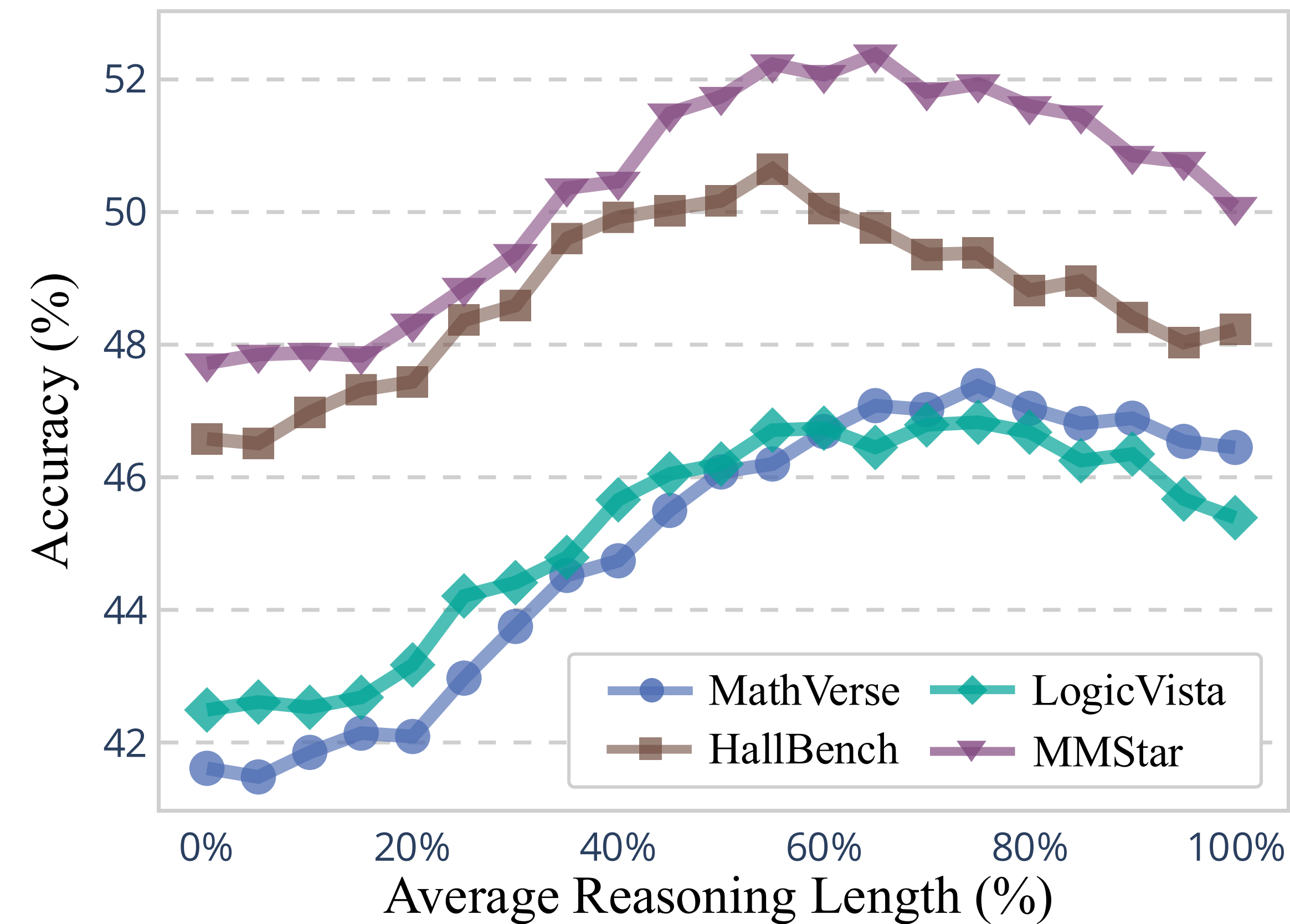
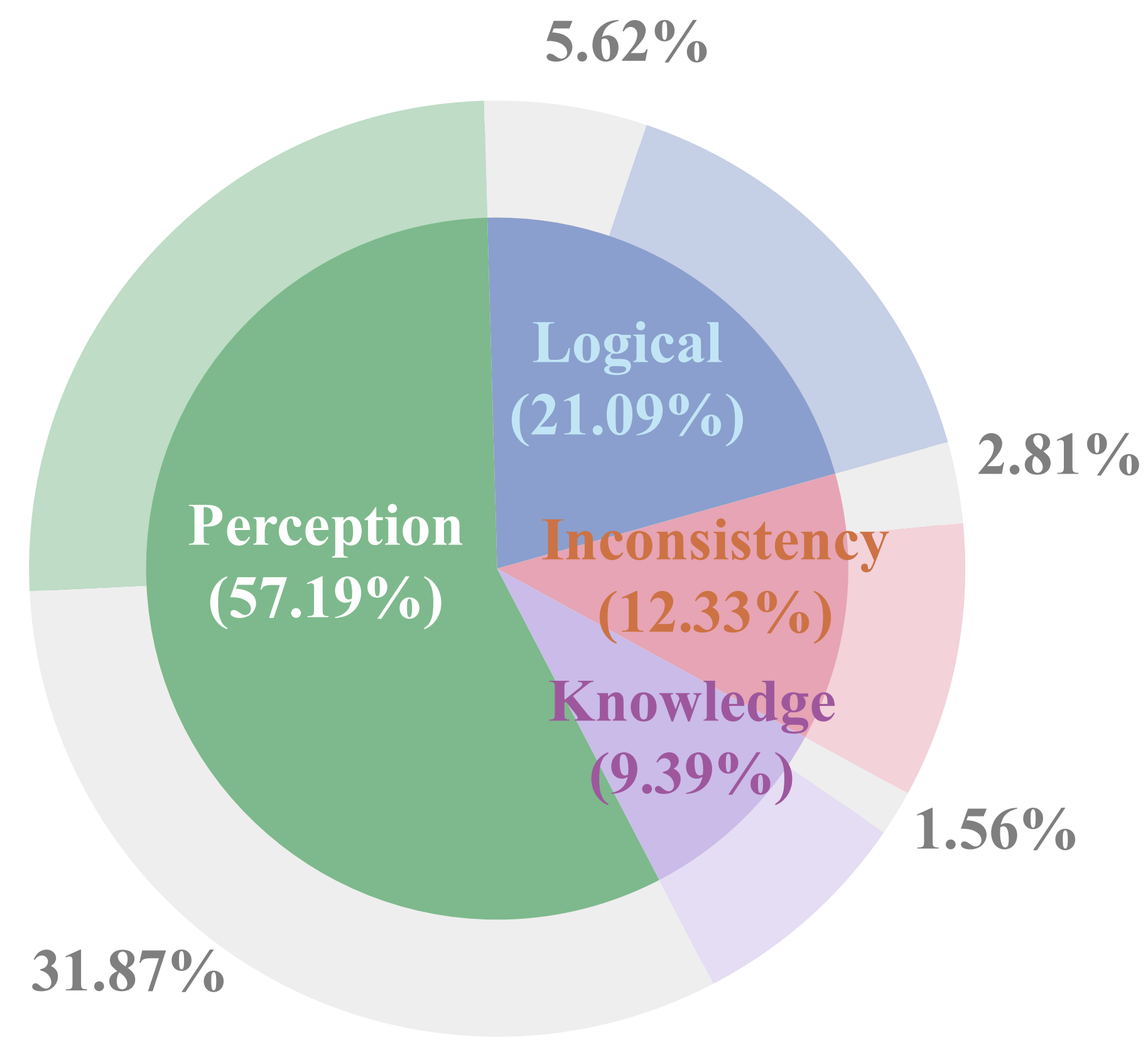
Question: What is the cause of this negative effect?

Dual Nature of Reasoning



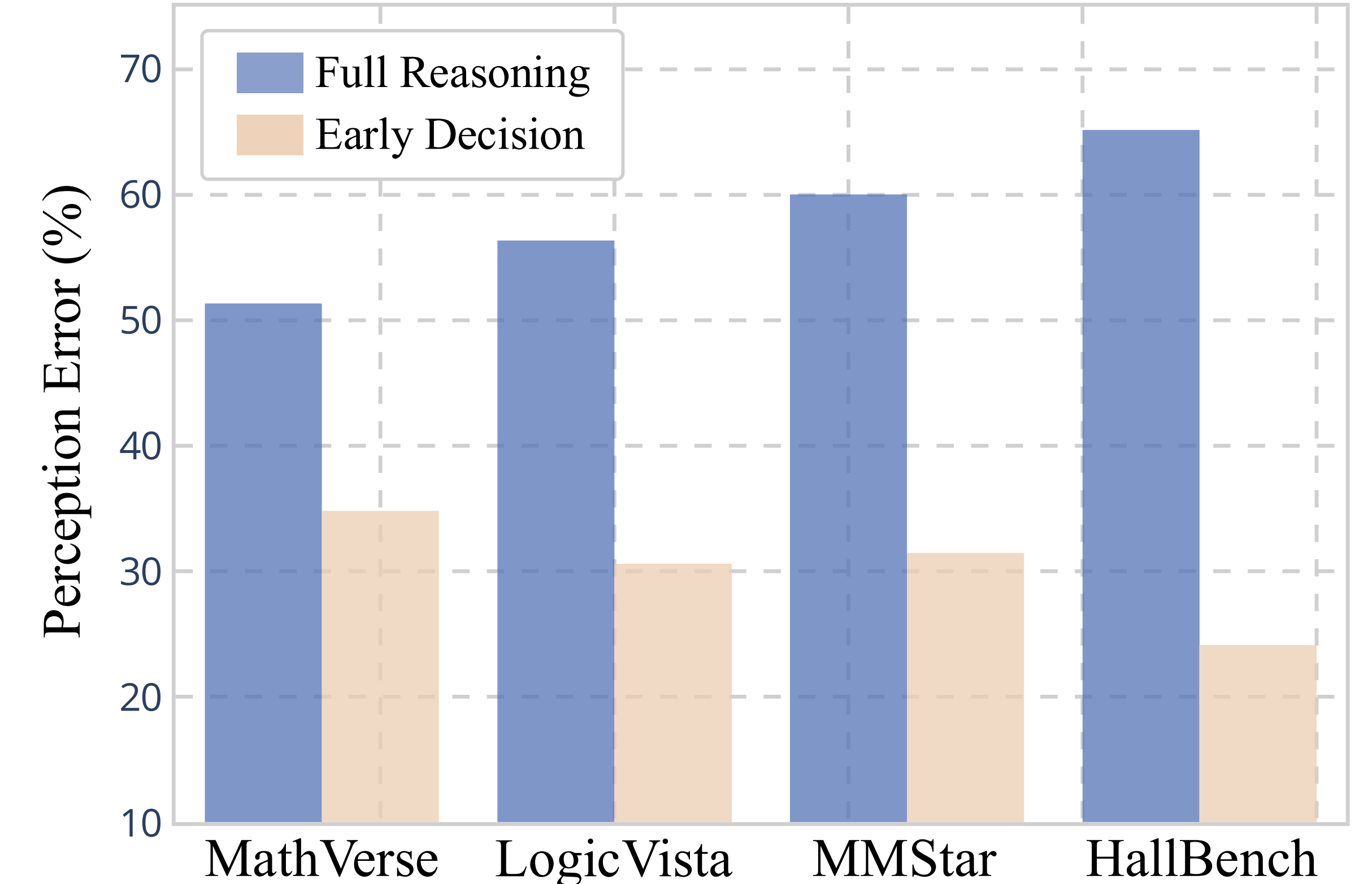
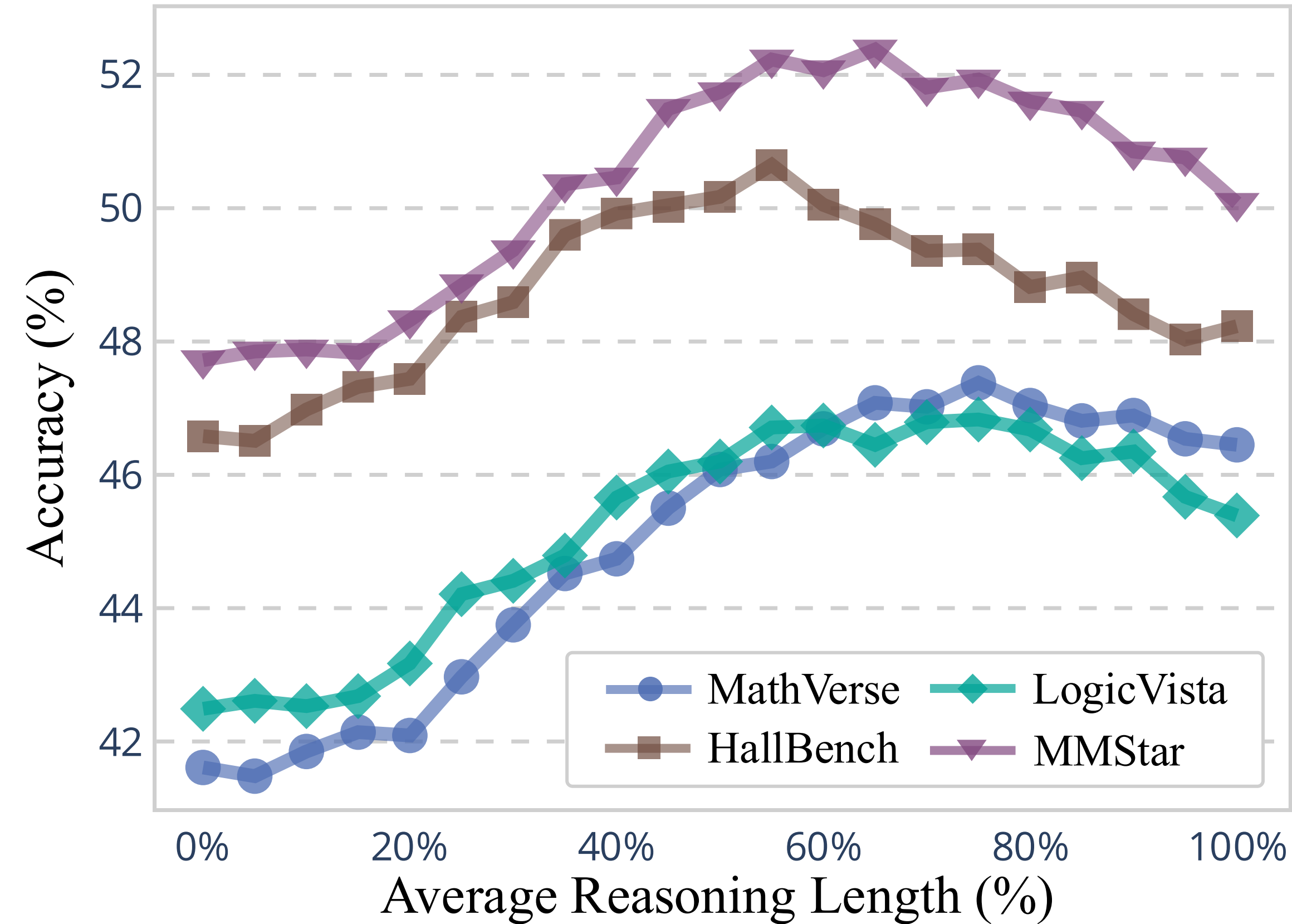
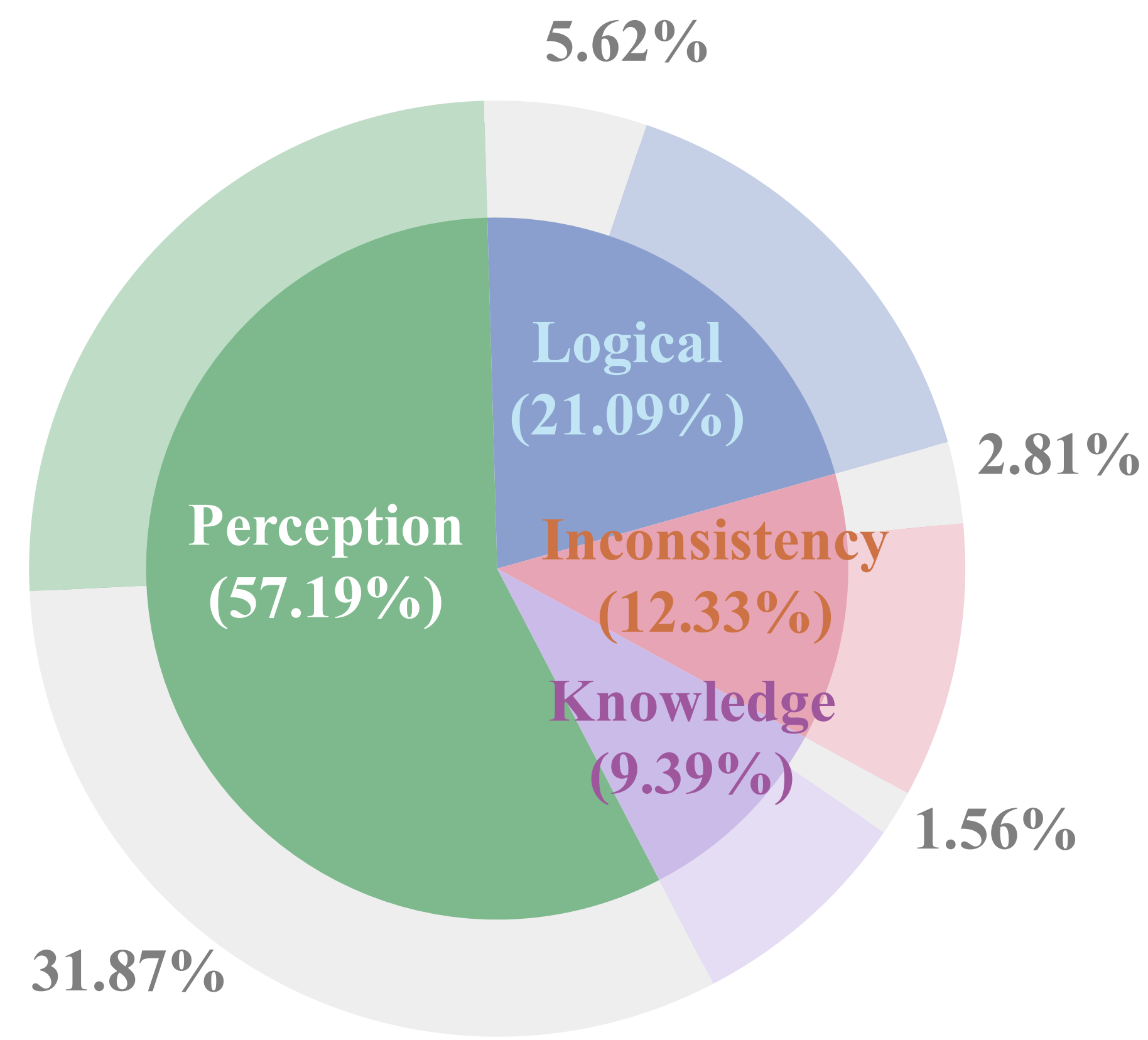
- We use GPT-4-mini to categorize the error cases under full reasoning into multiple types: Perception, Logical, Inconsistency and Knowledge.
 - Perception: errors come from the misunderstand regarding the image details.
 - Logical: the calculation/inference is incorrect, or reasoning does not make sense
 - Inconsistency: the conflict between reasoning process and final answer
 - Knowledge: the commonsense, or external knowledge of models is wrong
- We also measure the proportion of errors that can be remedied by early decision (gray colored in the pie chart)

Dual Nature of Reasoning



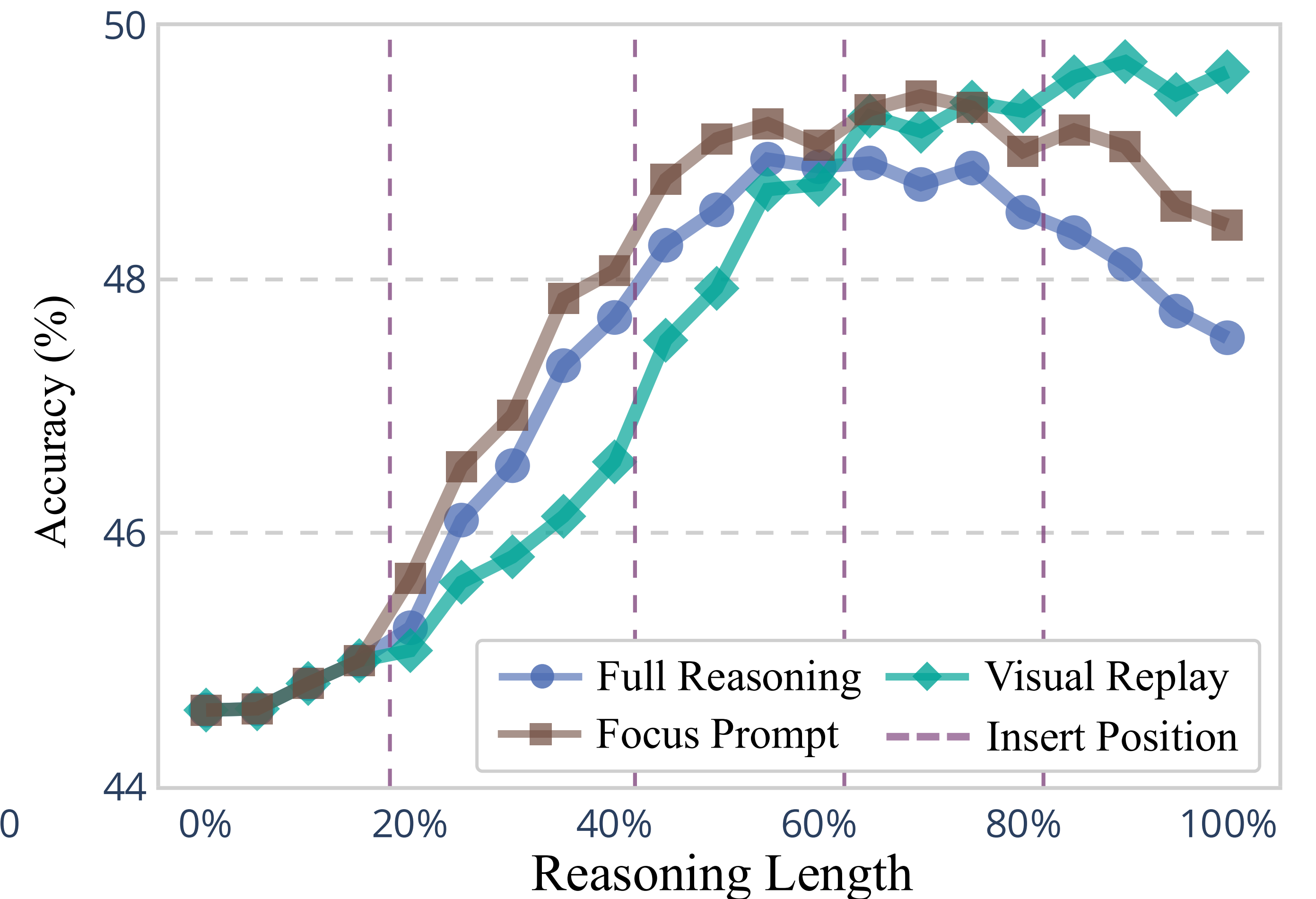
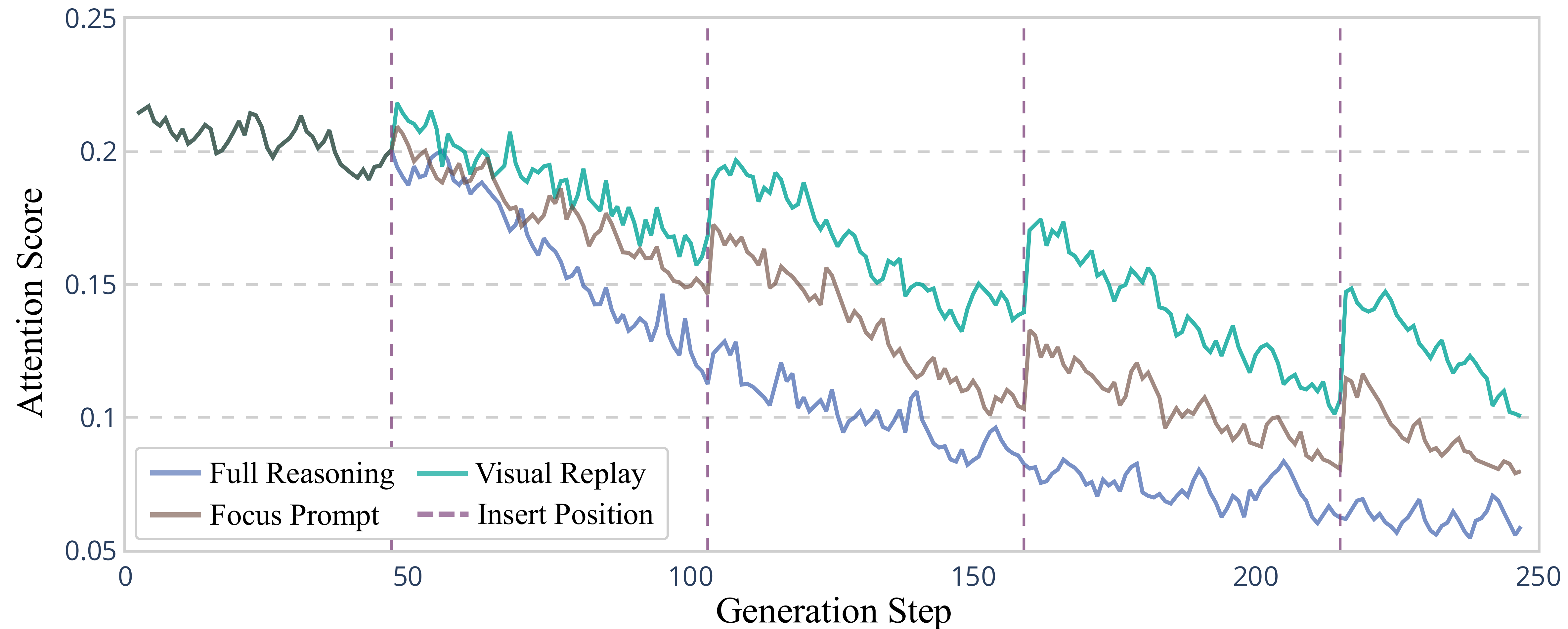
- **Most errors (57.19%) fall into perception errors, indicating current models, despite their strong reasoning capability, still exhibit weakness in basic image understanding.**
- **For this dominant type, VLMs can actually reveal the correct answers over half of the cases (31.87%) by early decision, implying that they are not incapable, but prolonged reasoning may instead compromise its perceptual ability.**
- **Overall, almost half of the error cases can be corrected by early decision (41.86%)**

Dual Nature of Reasoning



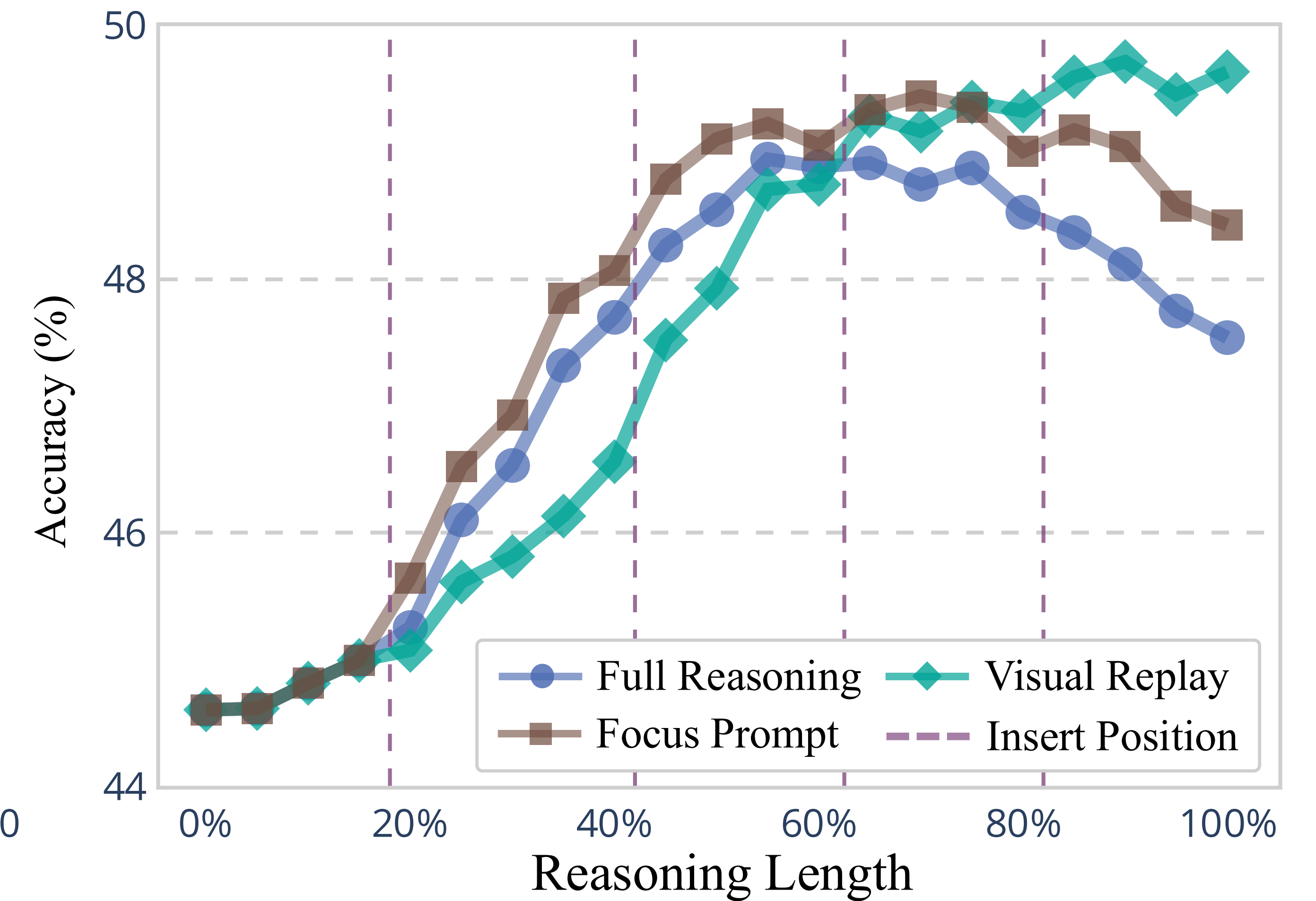
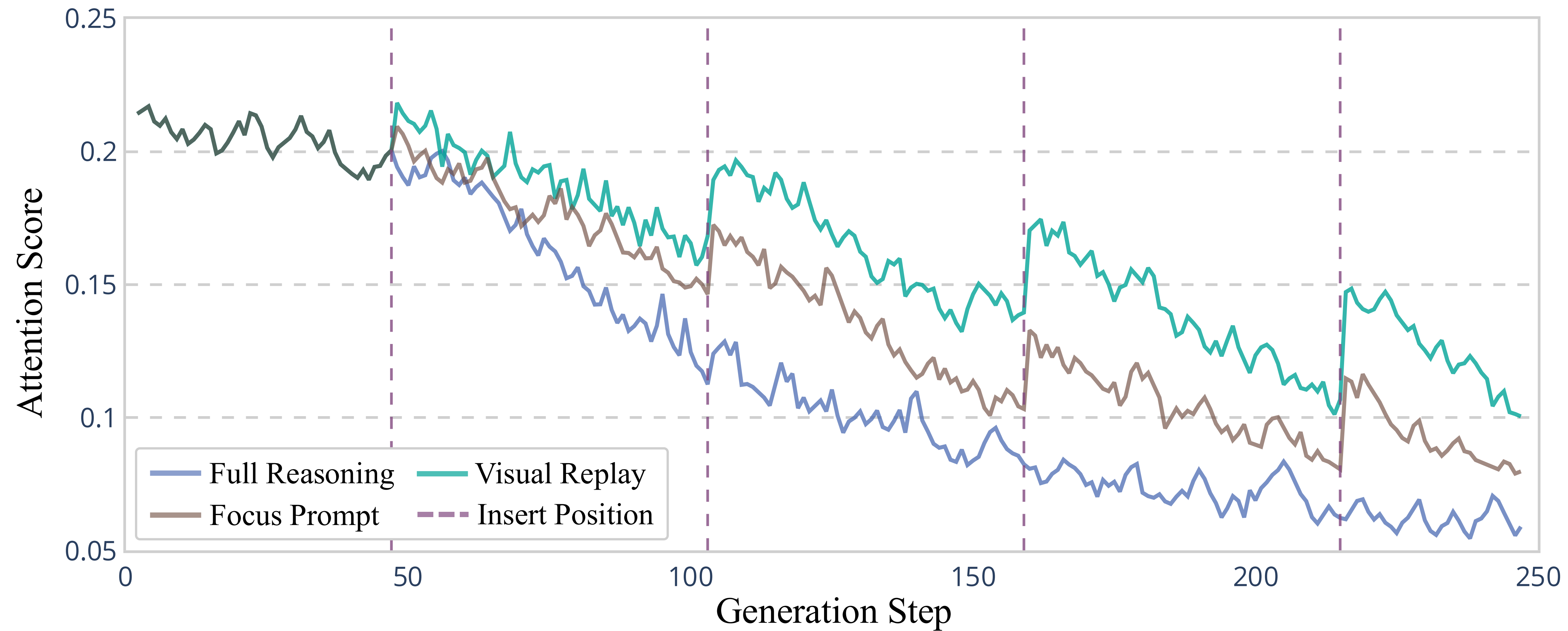
- **We also visualize the perception error rate of full reasoning and early decision on these benchmarks.**
- **We find that in visually dependent benchmarks, e.g., HallBench, perception errors account for the largest proportion, and shortening the reasoning length can reveal the correct answers in most cases.**
- **There is a trade-off between reasoning capability and perceptual ability.**

Dual Nature of Reasoning



- **We hypothesize that visual forgetting is the main causes for the cons of VL reasoning.**
- **We propose two explorative solutions to verify our hypothesis**
 - **Visual Replay: let VLMs see the original image every few reasoning steps**
 - **Focus Prompt: explicitly instruct VLMs to attend to the input image by prompts**
- **We visualize the attention of visual tokens for each generation step, and the accuracy change over the reasoning progress, as well as the intervention position**

Dual Nature of Reasoning



- **Each time an image or instruction is inserted, the model's attention to visual tokens increases to some extent, while attention continues to decline as reasoning progresses**
- **Corresponding to attention, each insertion immediately boosts the contribution of reasoning to the model accuracy, especially in the later stages of reasoning, and directly providing images is more effective than indirectly prompting VLMs to attend to it.**
- **This verifies our assumptions that the negative end of reasoning comes from forgetting.**

Vision-Anchored Policy Optimization

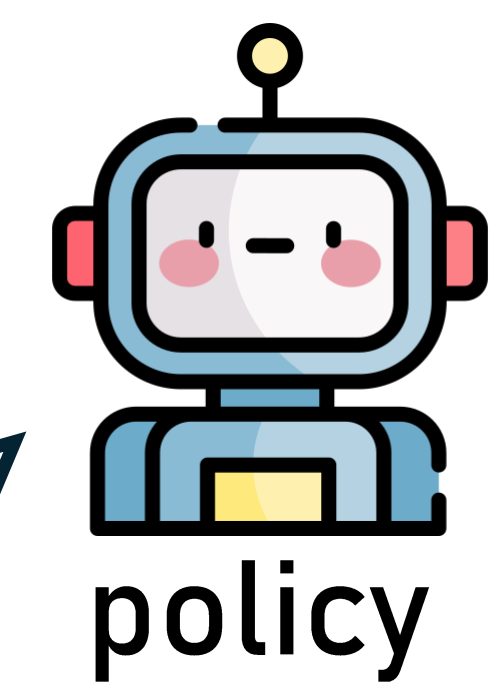
The current reasoning models suffer serious visual forgetting issues. We need to train a model where the reasoning process should be tied with visual information.



rollout

Q: What color is the rear wheel of the bus directly behind the small car?

<think>I am asked to identify the color in this photo, let's go through reasoning step by step:
 1. There is only one small yellow car in the middle right.
 <anchor>There are two buses facing left in the photo.
 Is this claim correct? Answer (yes/no): yes </anchor> ✓
 2. Behind the car, I find a motor, bicycle and two buses.
 3. The green bus is right in the back of the car.
 <anchor>The golden bike is behind the green motorcycle.
 Is this claim correct? Answer (yes/no): yes </anchor> ✗
 4. The rear wheel of the green bus presents blue.</think>
 <answer>The color is blue</answer>



policy

verify

- Accuracy
- Perception
- Format



GPT-4

generate

- the small car is yellow
 - the gray bus is facing left
 - ...
 - there are no green bicycles
 - a bus in front of the golden bike
- visual claims

insert anchors

$$R_{pcp} = \sum_j w_j R_j * R_{acc}$$

$$\sum_j w_j = 1, w_j < w_{j+1}$$

$$R_{acc} = \begin{cases} 1 & \text{if correct} \\ 0 & \text{otherwise} \end{cases}$$

$$R_{fmt} = \begin{cases} 1 & \text{if matched} \\ 0 & \text{otherwise} \end{cases}$$

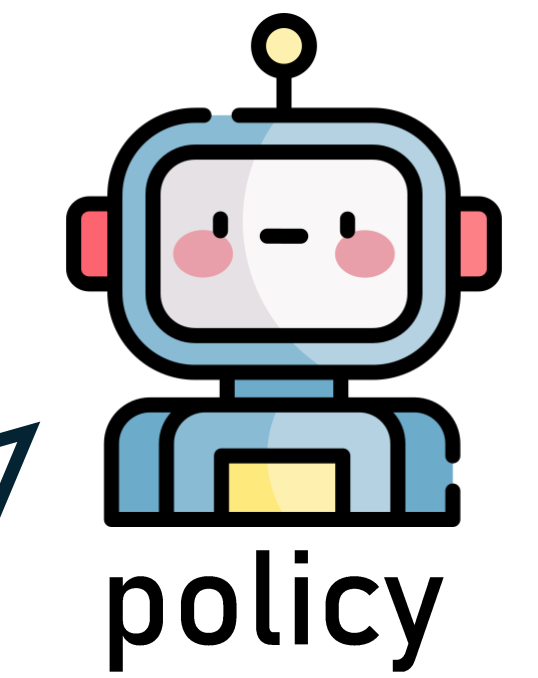
Vision-Anchored Policy Optimization



rollout

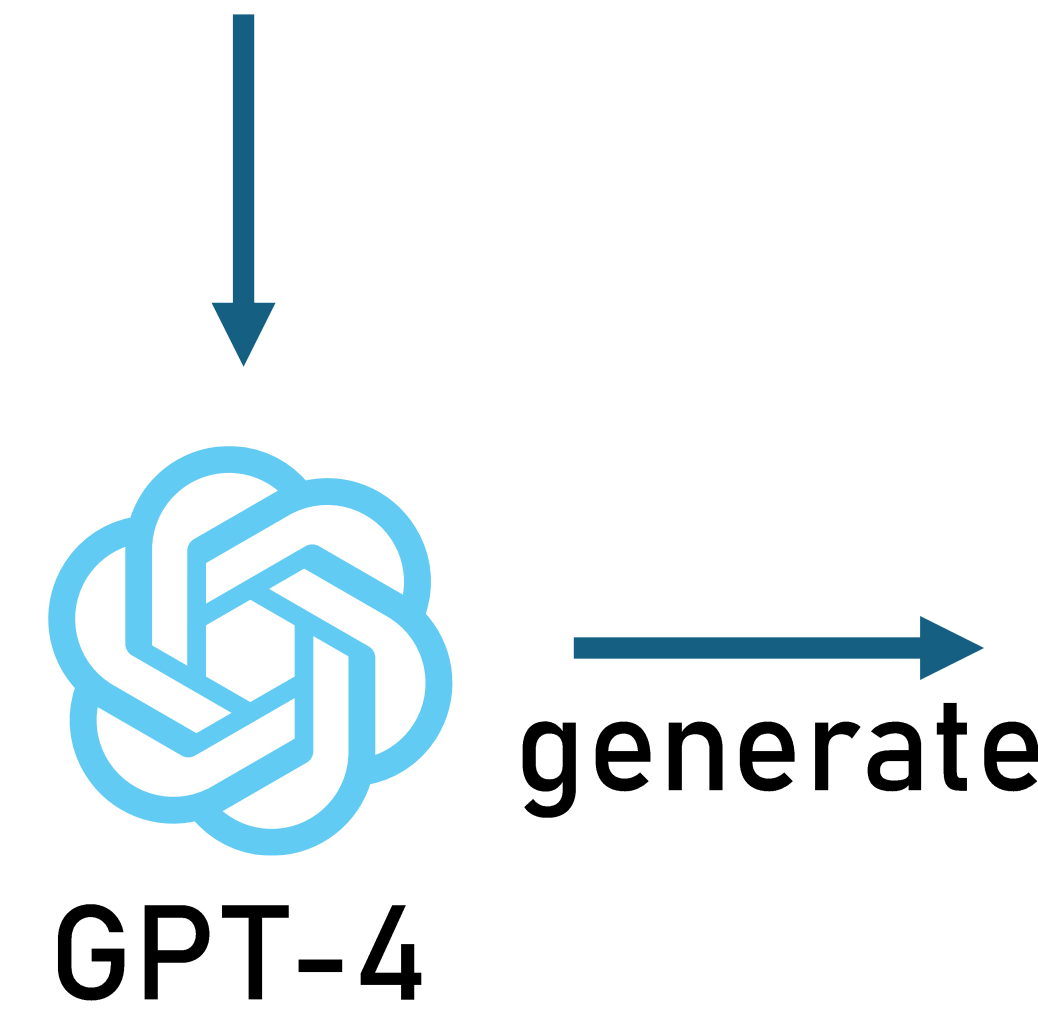
Q: What color is the rear wheel of the bus directly behind the small car?

<think>I am asked to identify the color in this photo, let's go through reasoning step by step:
 1. There is only one small yellow car in the middle right.
 <anchor>There are two buses facing left in the photo.
 Is this claim correct? Answer (yes/no): yes </anchor> ✓
 2. Behind the car, I find a motor, bicycle and two buses.
 3. The green bus is right in the back of the car.
 <anchor>The golden bike is behind the green motorcycle.
 Is this claim correct? Answer (yes/no): yes </anchor> ✗
 4. The rear wheel of the green bus presents blue.</think>
 <answer>The color is blue</answer>



verify

- Accuracy
- Perception
- Format



- the small car is yellow
 - the gray bus is facing left
 - ...
 - there are no green bicycles
 - a bus in front of the golden bike
- visual claims

insert anchors

$$R_{pcp} = \sum_j w_j R_j * R_{acc}$$

$$\sum_j w_j = 1, w_j < w_{j+1}$$

$$R_{acc} = \begin{cases} 1 & \text{if correct} \\ 0 & \text{otherwise} \end{cases}$$

$$R_{fmt} = \begin{cases} 1 & \text{if matched} \\ 0 & \text{otherwise} \end{cases}$$

- **VAPO:** we evaluate the perception capability of VLMs every few steps, and reward them if visual claims can be judged correctly

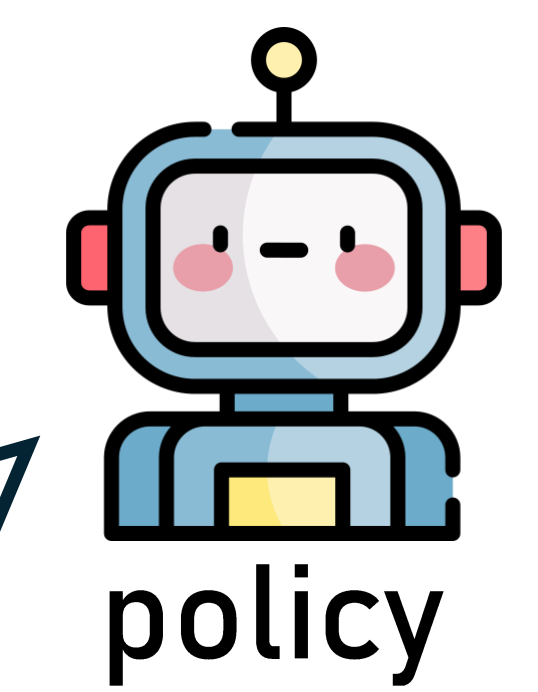
Vision-Anchored Policy Optimization



rollout

Q: What color is the rear wheel of the bus directly behind the small car?

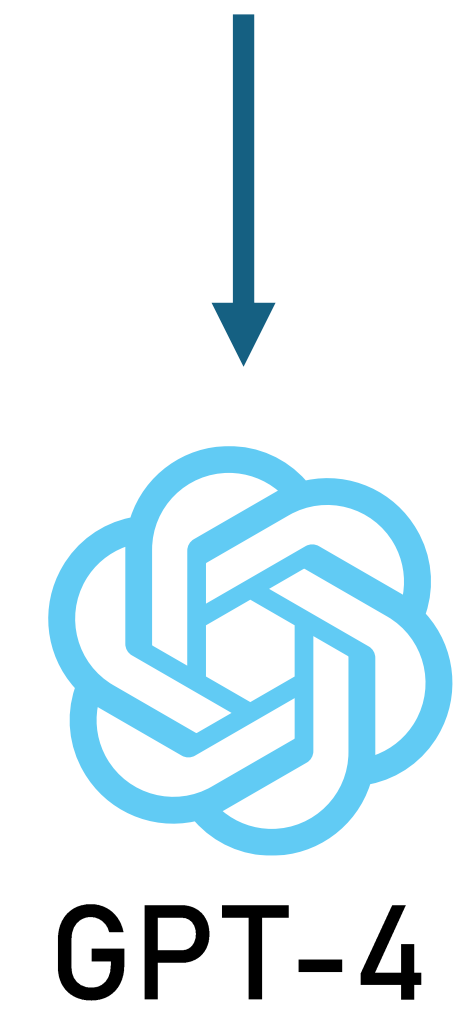
<think>I am asked to identify the color in this photo, let's go through reasoning step by step:
 1. There is only one small yellow car in the middle right.
 <anchor>There are two buses facing left in the photo.
 Is this claim correct? Answer (yes/no): yes </anchor> ✓
 2. Behind the car, I find a motor, bicycle and two buses.
 3. The green bus is right in the back of the car.
 <anchor>The golden bike is behind the green motorcycle.
 Is this claim correct? Answer (yes/no): yes </anchor> ✗
 4. The rear wheel of the green bus presents blue.</think>
 <answer>The color is blue</answer> \boxed{answer}



policy

verify

- Accuracy
- Perception
- Format



generate

- the small car is yellow
 - the gray bus is facing left
 - ...
 - there are no green bicycles
 - a bus in front of the golden bike
- visual claims

insert anchors

$$R_{pcp} = \sum_j w_j R_j * R_{acc}$$

$$\sum_j w_j = 1, w_j < w_{j+1}$$

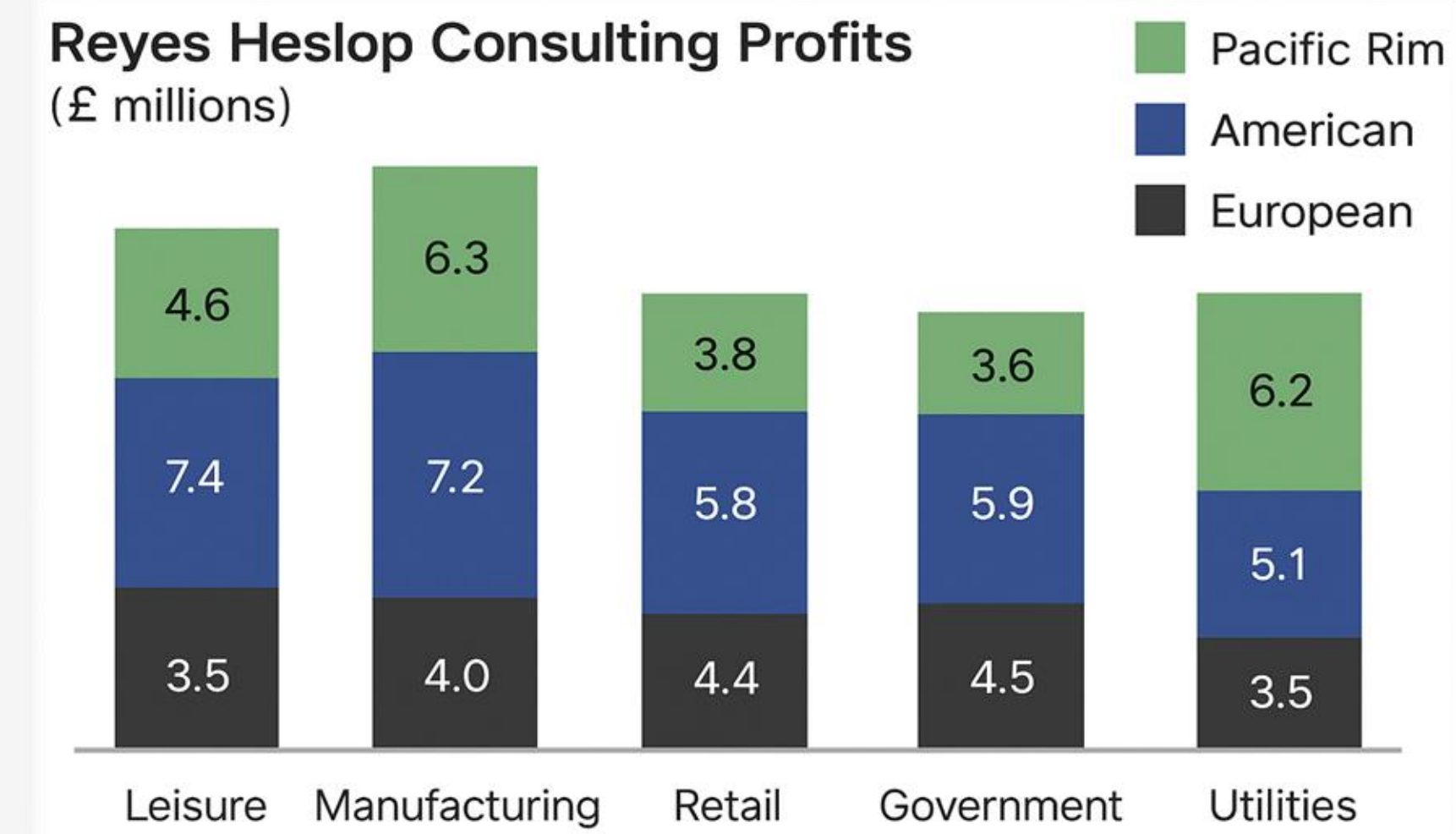
$$R_{acc} = \begin{cases} 1 & \text{if correct} \\ 0 & \text{otherwise} \end{cases}$$

$$R_{fmt} = \begin{cases} 1 & \text{if matched} \\ 0 & \text{otherwise} \end{cases}$$

• Intuitively, this is like dropping a visual anchor in the reasoning process to ensure VLMs are visually grounded during their prolonged rollout.

LogicVista (Chart)

Q: What is the total profit of top three industries in America? GT Answer: 20.5 DT Answer: 24.8



Early Decision

<think> We are given a chart from Reyes Heslop. Let's identify the profits of each industry:

1. Leisure: 4.6 (Pacific Rim), 7.4 (American) and 3.5 (European)

Time is up. I should stop thinking and give an answer based on what I have.</think>

<answer>The sum of the profit is 20.5 millions</answer>

Full Reasoning

<think> We are given a chart from Reyes Heslop. Let's identify the profits of each industry:

1. Leisure: 4.6 (Pacific Rim), 7.4 (American) and 3.5 (European)
2. Manufacturing: 6.3 (Pacific Rim), 7.2 (American) and 4.0 (European)
3. ...

4. Utilities: 6.2 (Pacific Rim), 8.1 (American) and 3.5 (European)

Back to the question, the top three industries are Leisure (7.4), Manufacturing (7.2), and Utilities (8.1). Let's sum the profit up: $7.4 + 7.2 + 8.1 = 22.7$ (millions).</think>

<answer>The total profit is 22.7 millions.</answer>

MMStar (Counting)

Q: In this given image, can you determine the number of white cats that appear? GT Answer: 4 DT Answer: 6



Early Decision

<think> Okay, we are asked how many cats are there in the photo. Let's have a look.

1. First, there is an adult white cat lying down with distinct blue and yellow eyes.
2. On the far left, a kitten lies on its back with tiny paws stretched upwards.

Time is up. I should stop thinking and give an answer based on what I have.</think>

<answer>There are 4 cats in this photo</answer>

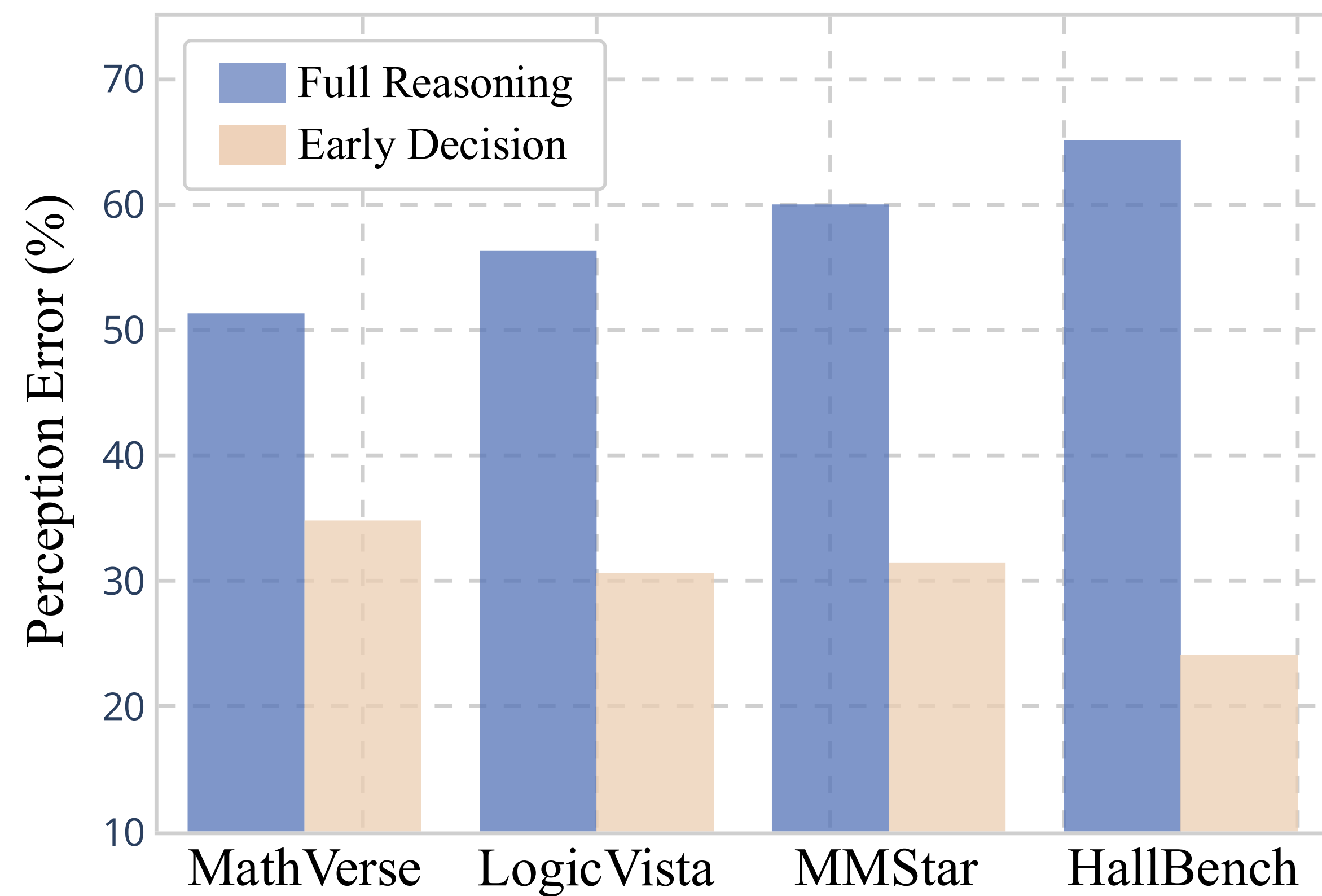
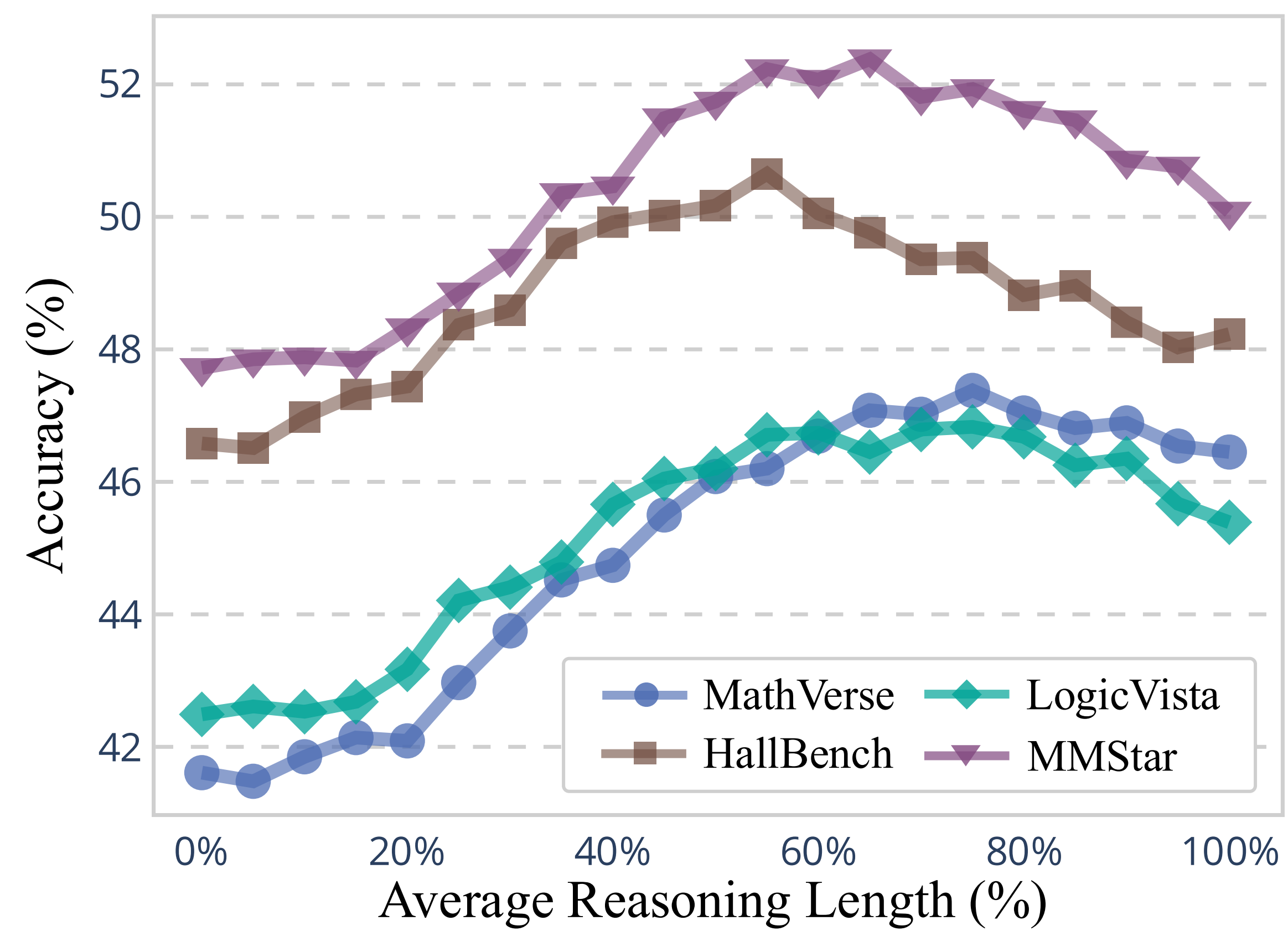
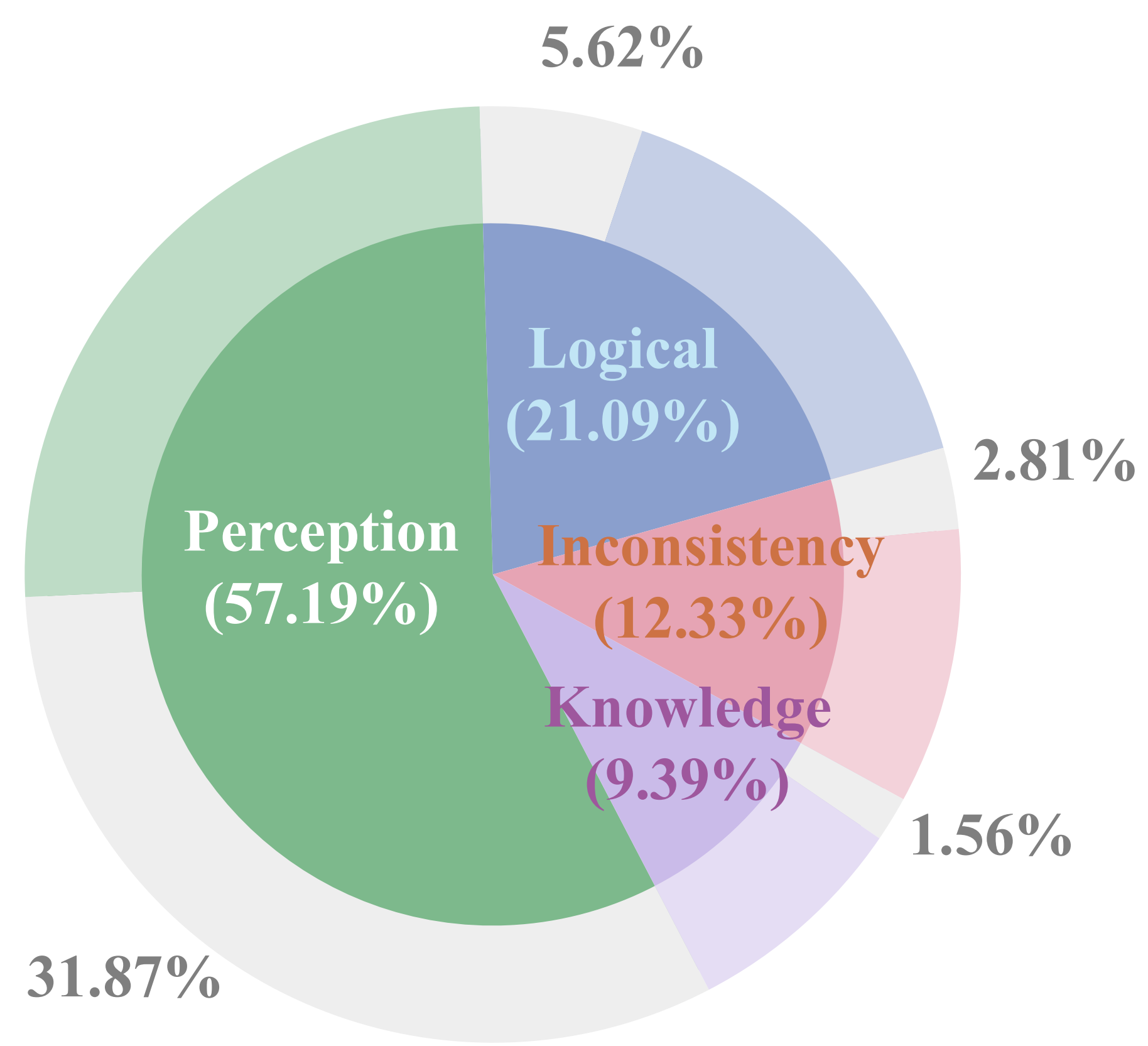
Full Reasoning

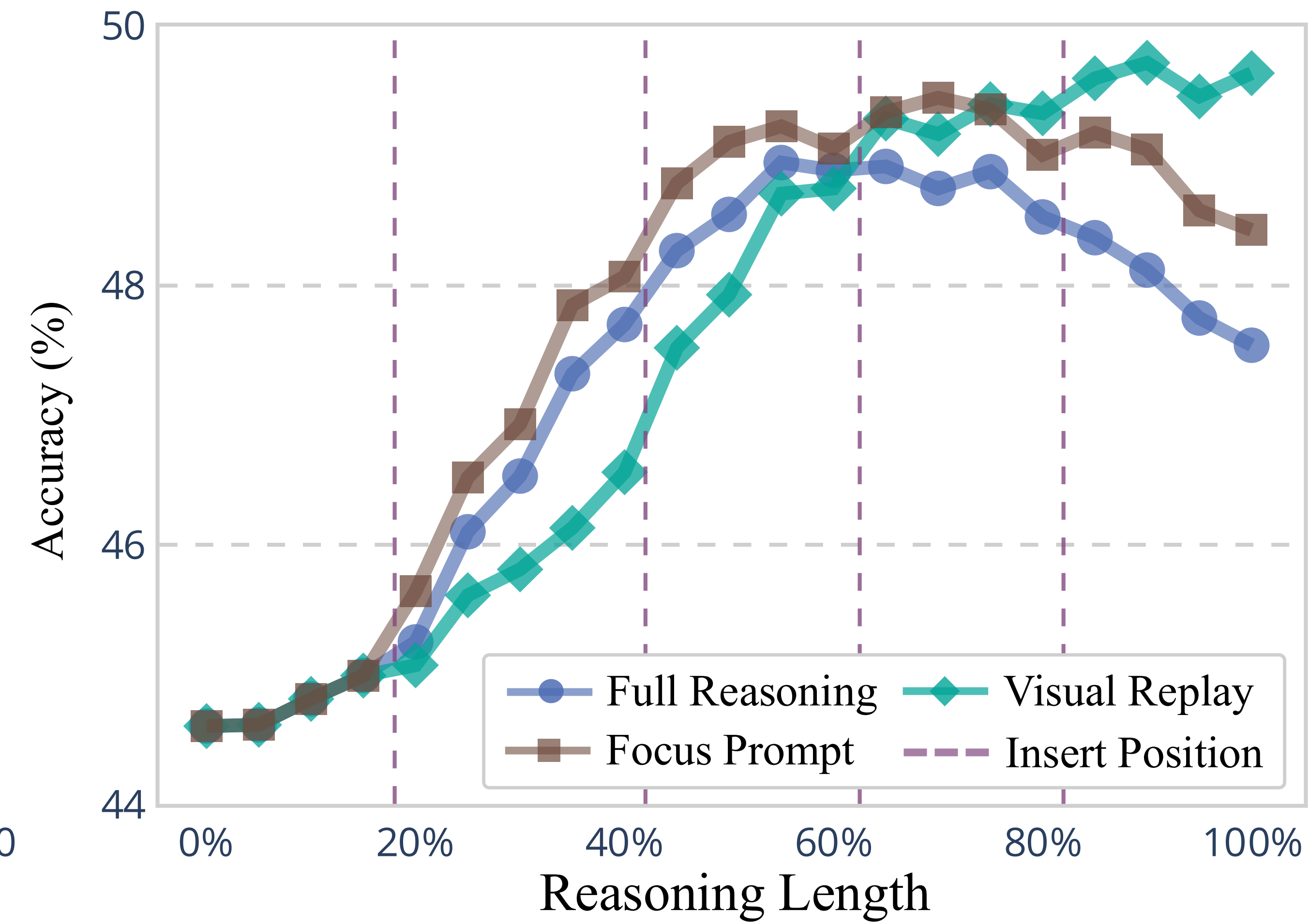
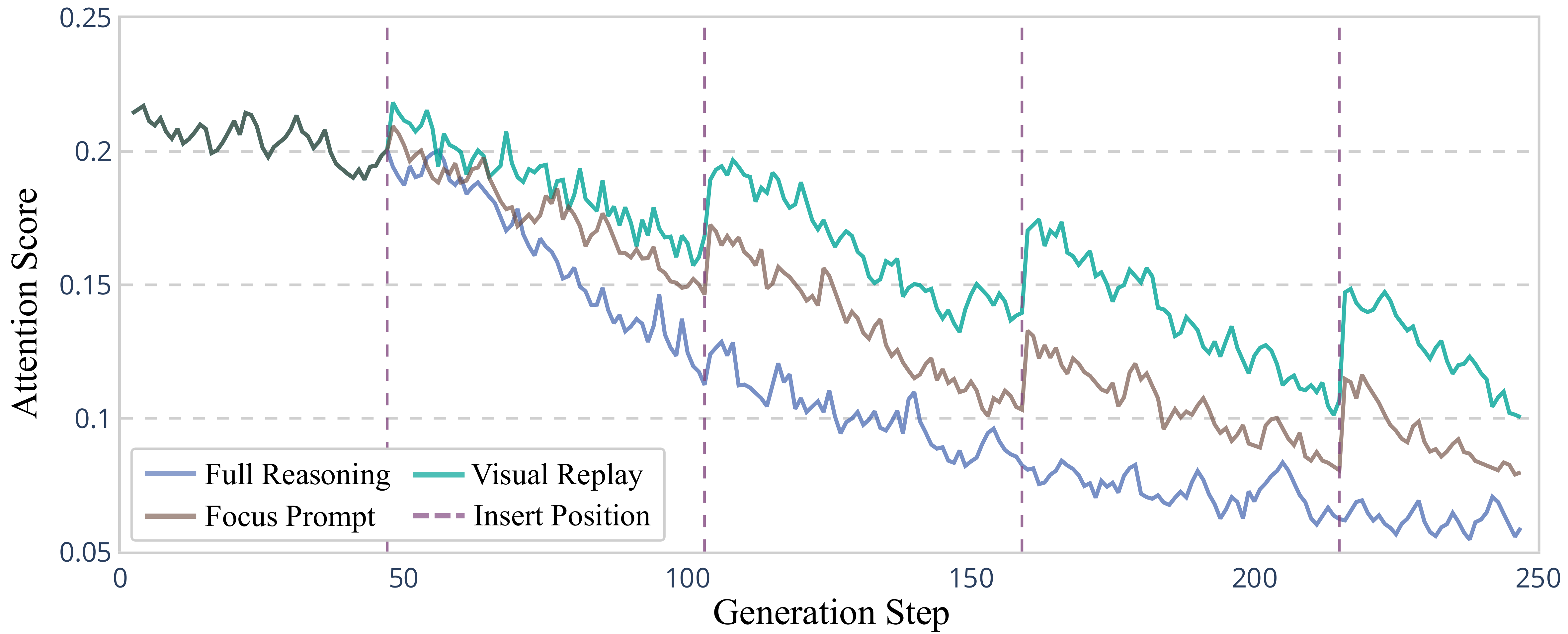
<think> Okay, we are asked how many cats are there in the photo. Let's have a look.

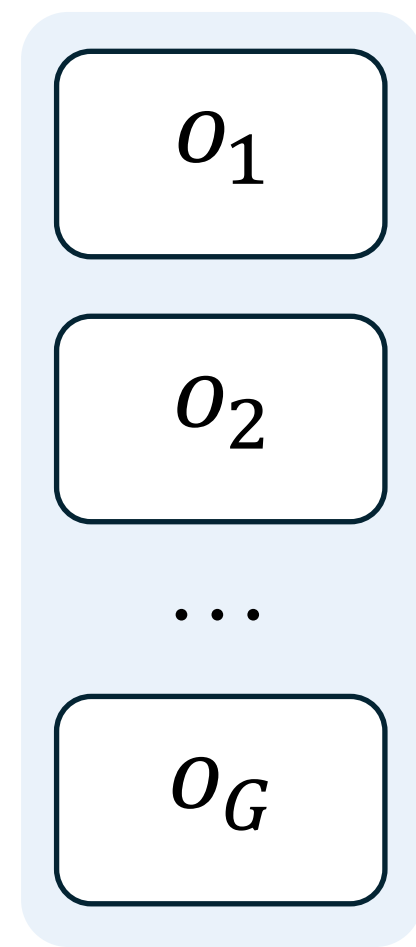
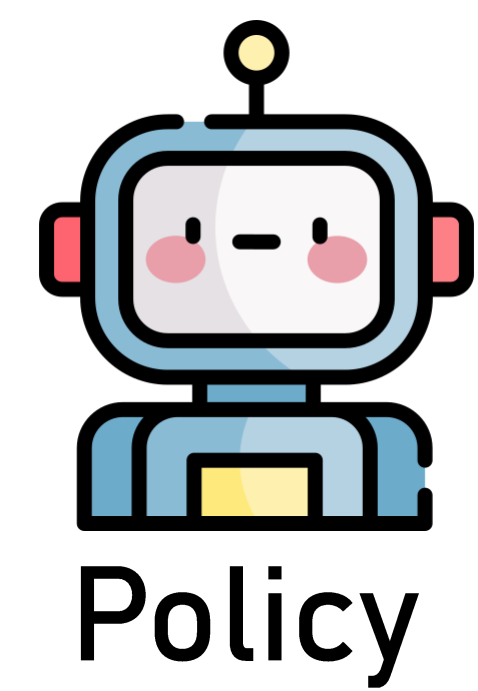
1. First, there is an adult white cat lying down with distinct blue and yellow eyes.
2. On the far left, a kitten lies on its back with tiny paws stretched upwards.
3. There is one cat in the middle-bottom of the photo, with head tilted to the side.
4. To the right, closest to the mother's head, a kitten is resting with belly showing.

5. On the far right, there is a kitten hidden beneath, with only its black eyes visible.

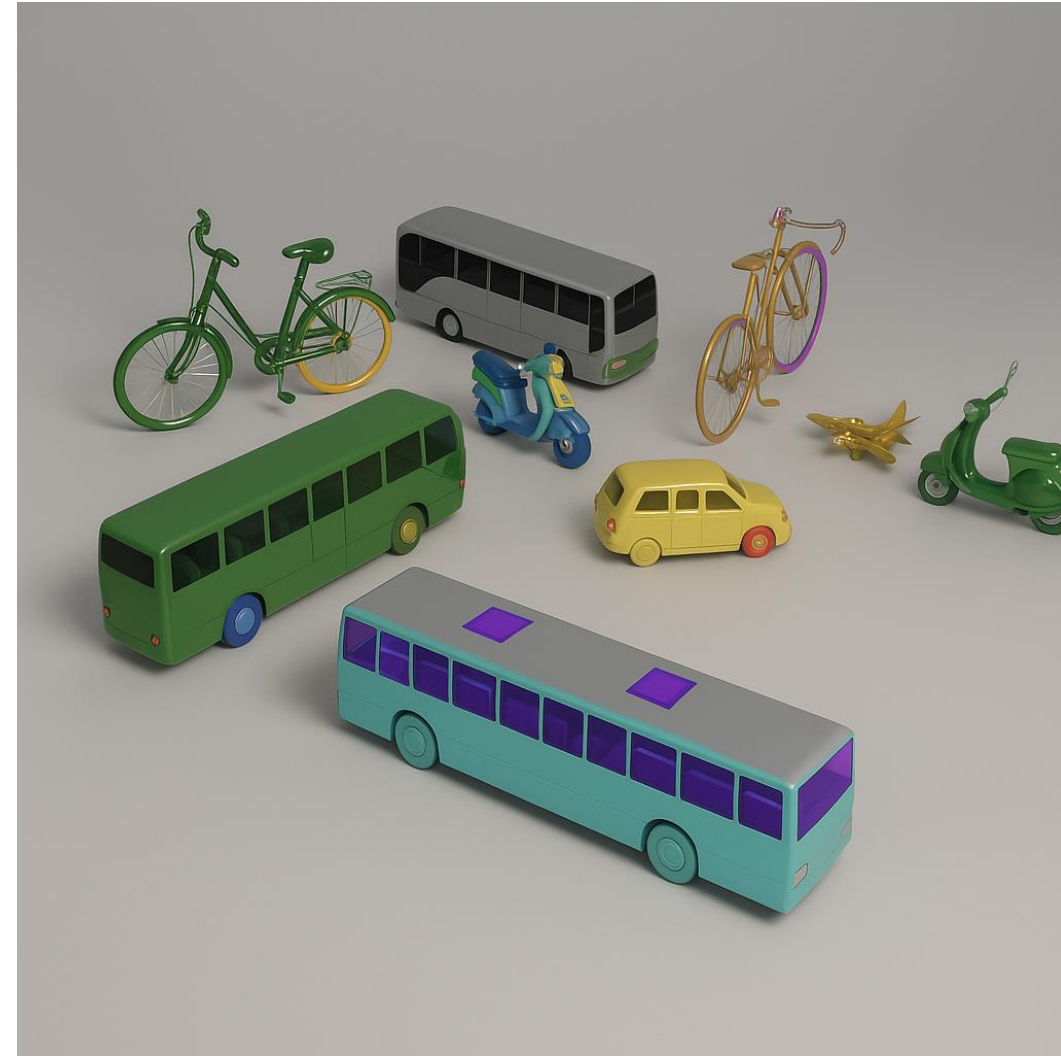
<answer>The number of cats is 5</answer>



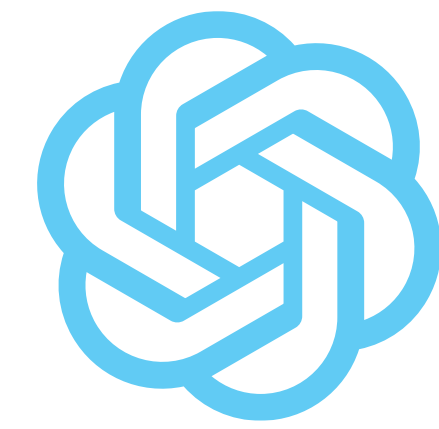




Q: What color is the rear wheel of the bus directly behind the small car?



input image



GPT-4



the small car is yellow

the gray bus is facing left

...

there are no green bicycles

a bus in front of the golden bike

proposed claims

