TRACE: A Framework for Analyzing and Enhancing Stepwise Reasoning in Vision-Language Models

Shima Imani* **Seungwhan Moon** **Lambert Mathias** **Lu Zhang** **Babak Damavandi**
Meta Reality Lab

Abstract

Reliable mathematical and scientific reasoning remains an open challenge for large vision–language models (VLMs). Standard final-answer evaluation often masks reasoning errors, allowing silent failures to persist. To address this gap, we introduce **TRACE** (*Transparent Reasoning And Consistency Evaluation*), a framework for analyzing, diagnosing, and improving reasoning in VLMs. At its core, TRACE leverages **Auxiliary Reasoning Sets (ARS)**, compact sub-question–answer pairs that decompose complex problems, evaluate intermediate steps through consistency-based metrics, and expose failures overlooked by standard evaluation. Our experiments show that consistency across ARS is linked to final-answer correctness and helps pinpoint the reasoning steps where failures arise, offering actionable signals for model improvement.

1 Introduction

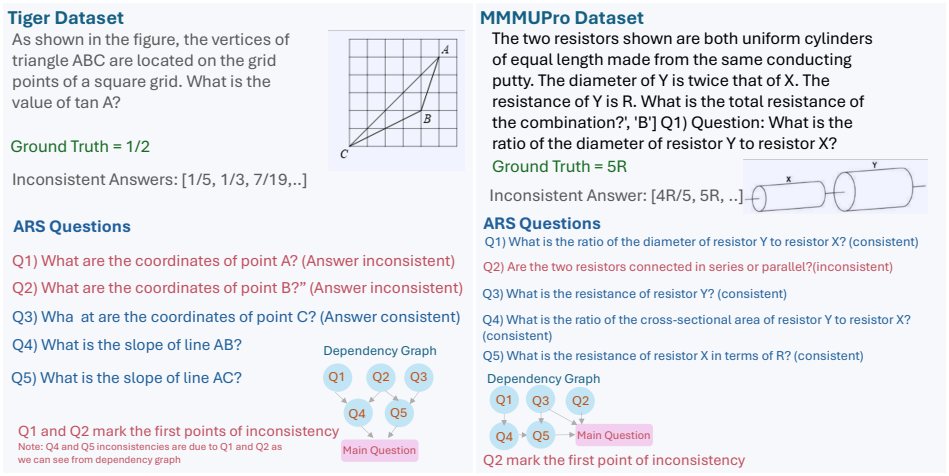

Figure 1: Illustration of the TRACE framework. (**Left**) A geometry question from TIGER-Lab/ViRL39K dataset TIGER-Lab [2025] with its ARS decomposition: Q1–Q3 extract coordinates, Q4–Q5 compute slopes, and the final question computes $\tan A$. Nodes highlighted in red indicate inconsistent answers across model generations, allowing us to localize the source of error. (**Right**) A physics problem from the MMMUPro dataset Yue et al. [2024] with its ARS decomposition. Nodes highlighted in red indicate inconsistent answers across model generations, helping to pinpoint where the reasoning fails.

*Corresponding author: shimaimani@meta.com

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MATH-AI.

Evaluating large vision-language models (VLMs) has predominantly focused on final-answer correctness. However, this metric often proves insufficient and misleading for comprehensive assessment. When a model yields an incorrect final answer, the precise step at which reasoning failed or how errors propagated through the computational graph remains obscure. Conversely, a correct final answer does not inherently guarantee a coherent or robust reasoning process; models can arrive at correct outcomes through flawed or inconsistent intermediate steps, masking underlying conceptual misunderstandings or accidental self-corrections. These limitations highlight critical, underexplored challenges: *How can we pinpoint failures in multi-step, multimodal VLM reasoning? What types of reasoning patterns produce "silent errors," where incorrect intermediate steps still lead to correct final answers? And to what extent does final-answer accuracy reliably reflect reasoning ability versus chance or memorization?*

To tackle these fundamental challenges, we introduce **TRACE** (*Transparent Reasoning And Consistency Evaluation*), a novel framework designed to enhance the diagnostic capabilities of VLM evaluation. TRACE systematically decomposes complex multimodal reasoning tasks into a series of interpretable, granular sub-questions, organized into **Auxiliary Reasoning Sets (ARS)**. By evaluating consistency across these intermediate answers, TRACE enables the precise localization of reasoning failures, providing granular, structured signals crucial for iterative model improvement. In contrast to conventional final-answer-centric evaluation, this decomposition not only reveals *whether* a model's prediction is correct, but critically, *how* it arrives at that prediction and identifies which specific reasoning steps are most susceptible to errors.

Building on TRACE, ARS maps sub-question dependencies into a reasoning graph to pinpoint the *first point of failure*. For example, Figure 1 shows how ARS tracks error propagation from coordinate identification in geometry or from inconsistencies in reasoning about resistor connections, even in cases where the final VLM answer does not fully reflect these intermediate errors.

Our contributions are:

1. We propose **TRACE**: A framework for transparent reasoning in multimodal models, breaking down complex tasks into Auxiliary Reasoning Sets (ARS).
2. A dependency-aware evaluation protocol that tracks consistency across intermediate steps to pinpoint the first point of failure.
3. A benchmark of 3.7k ARS question–answer pairs across 630 reasoning paths, aligned with STEM problems, for evaluating multimodal reasoning at both intermediate and final answer levels.
4. Experiments demonstrating TRACE uncovers reasoning errors missed by standard final-answer evaluation, enhancing interpretability, robustness, and evaluation quality.

2 Related Work

Multimodal reasoning benchmarks. Benchmarks like ScienceQA Lu et al. [2022], MathVista Lu et al. [2023], MMMU Yue et al. [2024], and MathVision Wang et al. [2024] evaluate VLMs on STEM tasks involving text and diagrams. While they reveal progress and limitations, they mainly assess final-answer accuracy, which cannot distinguish genuine reasoning from lucky guesses.

Intermediate reasoning and process supervision. Methods such as chain-of-thought prompting Wei et al. [2022], self-consistency Wang et al. [2022], and process supervision (e.g., GSM8K Cobbe et al. [2021]) improve performance and interpretability by incorporating intermediate steps. However, they often rely on natural language rationales, which are hard to evaluate automatically, or still depend on final-answer supervision, limiting error diagnosis in intermediate steps.

Problem decomposition and program-aided reasoning. Approaches like Program-of-Thoughts Chen et al. [2022], PAL Gao et al. [2022], and verifier-based pipelines Lightman et al. [2023] decompose reasoning into symbolic programs or use self-correction. While effective for arithmetic and robustness, they lack principled ways to diagnose failures at specific reasoning steps, especially in multimodal contexts.

3 Methodology

Building on these insights, we introduce **TRACE** a framework that structures and scrutinizes the intermediate reasoning steps of multimodal models. TRACE moves beyond conventional black-box mapping, where a model directly maps VLM(Question, Image) to a final answer. Instead, it introduces an intermediate **Auxiliary Reasoning Set (ARS)**,

$$\mathcal{S} = \{(q_i, a_i)\}_{i=1}^{n},$$

consisting of sub-question–answer pairs that explicitly capture intermediate reasoning steps. An ARS is defined such that each sub-question q_i paired with an answer a_i satisfies three properties:

- **Completeness:** The sub-questions collectively provide all information needed to solve the problem, with no redundancy. The ARS also extracts any required information from the image, so that the main question can be answered without directly referencing the figure.
- **Independence:** Each sub-question depends only on raw inputs or explicitly specified predecessors.
- **Soundness:** Sub-questions are answerable, non-overlapping, and do not leak the final answer.

The model then uses this structured set to produce its final answer, making the reasoning process transparent and enabling fine-grained evaluation of reasoning behavior.

Example. Consider a geometry problem illustrated in Figure 1, where the vertices of triangle ABC lie on a square grid and the task is to compute $\tan A$. The ARS decomposes the problem into sub-questions such as identifying the coordinates of points A, B, and C, and computing the slopes of lines AB and AC. This decomposition allows TRACE to analyze the model's reasoning process, track how intermediate answers propagate, and provide a structured view of the entire solution pathway.

Construction. ARS are generated using two complementary strategies:

- **Exploration:** Given the original question and a specialized prompt, the model generates diverse sub-questions along with their dependencies.
- **Exploitation:** Sub-questions are generated from candidate reasoning chains in two steps:
  1. *Step 1:* Given the question and image, the model produces a step-by-step reasoning answer.
  2. *Step 2:* Using the original question, the generated reasoning answer, and a prompt, the model generates sub-questions and their dependencies, ensuring the ARS captures the critical steps necessary to solve the problem.

Reasoning Graph. Each sub-question is annotated with metadata specifying its dependencies on other questions, text, or image inputs. Together, these dependencies form a directed acyclic reasoning graph (DAG) that encodes the structure of the reasoning process and defines the order in which the model answers sub-questions. Intermediate answers propagate forward through the DAG, and the final answer is computed based on these intermediate results. This structured execution allows TRACE to pinpoint exactly which reasoning steps lead to errors, enabling transparent analysis of the model's reasoning process.

3.1 Evaluation Metrics

TRACE assesses reasoning performance at both the final-answer and intermediate-step levels, providing a multi-faceted evaluation of model behavior. We adopt the following key metrics:

Consistency Score: Measures stability by checking whether the model produces identical answers to the same sub-question across multiple reasoning paths. We introduce a family of **consistency-based metrics** (Appendix A) that capture finer-grained stability at both path and question levels:

- *Path mean, deviation, and z-score consistency* capture stability within a reasoning trajectory.
- *Global mean consistency* captures stability across all paths for a question.

- *Consistency gap* compares path-level stability against the global question-level baseline.

First Failure Step (FFS): We define the **First Failure Step (FFS)** as the earliest sub-question in a reasoning path that ultimately yields an *incorrect* final answer (or one that deviates from the majority outcome when no gold answer is available), at which the path's response first diverges from the consensus established by other consistent paths. The FFS marks the initiation of a reasoning trajectory that leads away from the reliable solution, providing a diagnostic signal for how local deviations accumulate into final errors.

Together, these metrics provide a fine-grained and interpretable evaluation of reasoning behavior, enabling analysis of both the reliability and robustness of VLM reasoning at intermediate steps and at the final-answer level.

4 Experiments and Results

We evaluate TRACE on two challenging STEM benchmarks: the multimodal MMMUPro dataset (covering Physics, Math, and Chemistry) and a verifiable subset of the TIGER-Lab/ViRL39K dataset (500 Questions). Both datasets require multi-step reasoning grounded in visual or symbolic information, making them ideal for assessing robustness. For each question, we generate ARS sub-questions using `Llama-4-Maverick-17B-128E-Instruct` and sample multiple reasoning paths for each ARS set across different models, enabling TRACE to systematically analyze the consistency and propagation of intermediate reasoning steps within the structured ARS framework.

Our results confirm that consistency scores are correlated with correctness, enabling us to distinguish reliable from unreliable reasoning paths. Figure 2 presents scatter plots of path mean consistency versus global mean consistency for MMMUPro, highlighting a clearly defined "red zone" of low-consistency trajectories. This red zone comprises 4.5% of paths for `GPT-4.1`, 5.3% for `Llama-4-Maverick`, and 9.7% for `Qwen2.5-VL-72B`, yet accounts for the vast majority of incorrect predictions (87.9%, 76.9%, and 77.5%, respectively). A similar pattern is observed for `Llama-4-Maverick` on the TIGER-Lab/ViRL39K dataset, confirming the consistency of this signal across domains. Applying the First Failure Step (FFS) analysis further pinpoints the sub-question where reasoning first diverges, illustrating how small local mistakes can cascade into globally incorrect outcomes. A few examples are provided in Appendix A.4. Together, these findings establish TRACE as a diagnostic tool that not only predicts confidence but also localizes sources of error, offering actionable insights to improve the trustworthiness of model reasoning. Detailed results and breakdown analyses are provided in Appendix A.

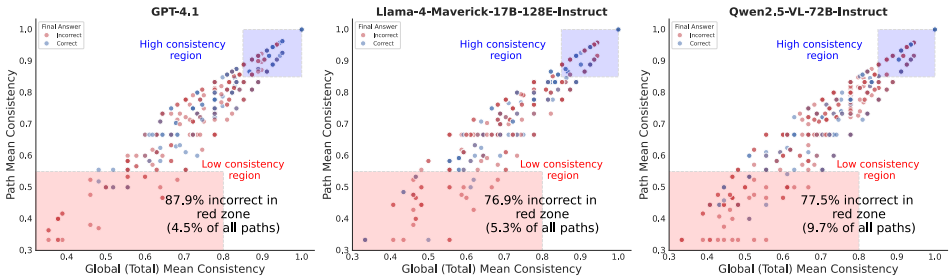

Figure 2: Scatter plots showing the relationship between path mean consistency and global (total) mean consistency for different models for the MMMUPro dataset. The red-shaded regions highlight low-consistency zones where a high fraction of paths result in incorrect answers.

5 Conclusion and Future Work

We introduced **TRACE**, a framework for transparent reasoning in multimodal models, which decomposes complex problems into stepwise reasoning traces. As part of TRACE, we provide a rigorously curated benchmark that evaluates models beyond final-answer accuracy, enabling fine-grained analysis of reasoning failures. Future work includes (i) training models to autonomously generate reliable reasoning traces, (ii) extending the benchmark to a broader set of multimodal tasks, and (iii) leveraging reasoning traces as structured rewards in RL methods such as GRPO to reduce reward hacking.

References

Wenhu Chen et al. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*, 2022.

Karl Cobbe et al. Training verifiers to solve math word problems (gsm8k). *arXiv preprint arXiv:2110.14168*, 2021.

Luyu Gao et al. Pal: Program-aided language models. *arXiv preprint arXiv:2211.10435*, 2022.

Danny Lightman et al. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Pan Lu, Swaroop Mishra, Tony Xia, et al. Learn to explain: Multimodal reasoning via thought chains for science question answering. *NeurIPS*, 2022.

Pan Lu, Hritik Bansal, et al. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

TIGER-Lab. Virl39k: A curated collection of 38,870 verifiable vision-language qa pairs. `https://huggingface.co/datasets/TIGER-Lab/ViRL39K`, 2025. Accessed: 2025-09-11.

Kun Wang et al. Measuring multimodal mathematical reasoning with math-vision dataset. *arXiv preprint arXiv:2402.14804*, 2024.

Xuezhi Wang et al. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Jason Wei et al. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

Xiang Yue et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark. *CVPR*, 2024.

A Consistency Metrics for Reasoning Paths

To evaluate the reliability of the reasoning process in TRACE, we introduce a set of consistency-based metrics that quantify the stability of reasoning paths across sub-questions and across sampled paths. These metrics help us understand whether correct answers are associated with more stable and coherent reasoning.

A.1 Definitions and Notation

Let a question be decomposed into an Auxiliary Reasoning Set (ARS):

$$\mathcal{S} = \{(q_i, a_i)\}_{i=1}^{n},$$

where each q_i is a sub-question and a_i is the corresponding answer. For each question, we sample K reasoning paths, each independently answering all sub-questions and producing a final answer.

Let:

- $A_{i,j}$: Answer to sub-question q_i in reasoning path j
- $C_{i,j}$: Agreement count for $A_{i,j}$ (i.e., how many paths gave the same answer to q_i as path j)
- P_j: The j-th reasoning path

We define the following metrics:

- **Path Mean Consistency** ($\texttt{path\_mean\_consistency}_j$): The average agreement across all sub-questions in path P_j:

$$\texttt{path\_mean\_consistency}_j = \frac{1}{n} \sum_{i=1}^{n} C_{i,j}$$

- **Path Deviation Consistency** ($\texttt{path\_deviation\_consistency}_j$): The standard deviation of agreement scores across sub-questions in path P_j:

$$\texttt{path\_deviation\_consistency}_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(C_{i,j} - \texttt{path\_mean\_consistency}_j)^2}$$

- **Path Z-score Consistency** ($\texttt{path\_zscore\_consistency}_j$): A normalized version of the path consistency:

$$\texttt{path\_zscore\_consistency}_j = \log\left(\frac{\sum_{i=1}^{n} C_{i,j} - \texttt{path\_mean\_consistency}_j + \epsilon}{\texttt{path\_deviation\_consistency}_j + \epsilon}\right)$$

- **Global (Total) Mean Consistency** ($\texttt{total\_mean\_consistency}$): The average agreement across all sub-questions and all paths for a given question:

$$\texttt{total\_mean\_consistency} = \frac{1}{nK}\sum_{i=1}^{n}\sum_{j=1}^{K} C_{i,j}$$

- **Consistency Gap** ($\texttt{consistency\_gap}_j$): The difference between a path's consistency and the global average:

$$\texttt{consistency\_gap}_j = \texttt{path\_mean\_consistency}_j - \texttt{total\_mean\_consistency}$$

A.2 Running Example

Consider the following question:

> *The object on the left is x grams, and the object on the right is 50 grams. Represent the following quantitative relationship with an inequality.*

The ARS for this question includes the following sub-questions:

- Q1: Is the scale balanced?
- Q2: Which side of the scale is lower?
- Q3: What is the mass of the object on the right side of the scale?
- Q4: What inequality represents the relationship? (Main question)

Suppose we sample $K = 5$ reasoning paths. For each sub-question, we compute how many paths gave the same answer. This gives us a consistency matrix $C \in \mathbb{R}^{n\times K}$.

Let's say for path P_3, the agreement counts across sub-questions are:

$$C_{\cdot,3} = [4, 5, 5, 4]$$

Then:

$$\texttt{path\_mean\_consistency}_3 = \frac{4+5+5+4}{4} = 4.5$$

Assume the global mean consistency for this question is:

$$\texttt{total\_mean\_consistency} = 4.2$$

Then the consistency gap for path P_3 is:

$$\texttt{consistency\_gap}_3 = 4.5 - 4.2 = +0.3$$

This means that path P_3 is more consistent than the average path for this question.

A.3 Empirical Results

We analyze the relationship between consistency metrics and correctness.

Model	Dataset	Path Mean Consistency	Path Z-score Consistency	Final Correctness
Llama-4-Maverick-17B-128E-Instruct	TIGER-Lab/ViRL39K	0.79 0.92	3.83 5.83	Incorrect Correct
GPT-4.1	MMMUPro	0.787 0.907	3.77 5.73	Incorrect Correct
Qwen2.5-VL-72B-Instruct	MMMUPro	0.772 0.853	3.98 4.88	Incorrect Correct
Llama-4-Maverick-17B-128E-Instruct	MMMUPro	0.806 0.903	4.38 5.62	Incorrect Correct

Table 1: Comparison of models on path consistency metrics and final answer correctness across datasets. Each row pair shows the performance split by answer correctness.

Correctness vs. Path Consistency. Our results show a clear correlation between consistency and correctness. Reasoning paths that ultimately produce the correct final answer exhibit substantially higher stability across sub-questions. In particular, correct paths achieve both higher mean consistency and higher z-score consistency compared to incorrect ones:

These differences show that agreement across sub-questions carries predictive value, with more consistent trajectories being substantially more likely to yield correct outcomes. This finding supports our central claim that focusing solely on final answers overlooks important indicators of reliability, whereas sub-question consistency provides a complementary lens for assessing reasoning quality.

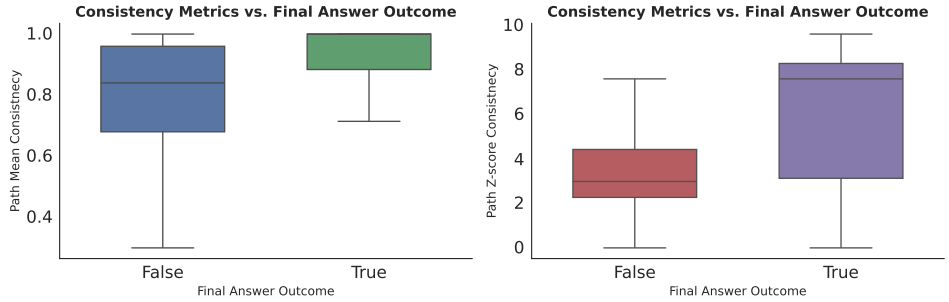

Figure 3: Path consistency metrics by final answer correctness, Llama-4-Maverick-17B-128E-Instruct results on Tiger dataset. Correct trajectories show markedly higher consistency, highlighting its predictive value for reasoning reliability.

We further investigate whether path-level consistency can serve as a confidence signal. Focusing on the TIGER-Lab/ViRL39K dataset with the `Llama-4-Maverick` model, we observe that mean path consistency is positively correlated with final answer correctness ($r = 0.362$, $p = 3.22 \times 10^{-76}$). By thresholding on mean path consistency, the accuracy of retained paths increases monotonically with the threshold: paths with perfect consistency (≥ 1.0) achieve 71% accuracy, compared to 50% overall. Similar patterns hold on the MMMUPro dataset across all evaluated models, indicating that consistency is a robust predictor of reasoning reliability and can guide the selection or reranking of reasoning paths.

We further examine the relationship between path-level and question-level consistency on the TIGER-Lab/ViRL39K dataset using `Llama-4-Maverick`. Figure 4 shows a scatter plot of path mean consistency versus global (total) mean consistency, with points colored by final answer correctness. Correct paths (blue) cluster in the top-right region of the plot, where both path-level and global consistency are high. However, some incorrect paths (red) also appear in this region, indicating that high consistency alone does not guarantee correctness. In contrast, incorrect paths are concentrated in the bottom-left, where both local and global consistency are low, while the middle region contains a mix of correct and incorrect paths. These patterns suggest that consistency provides a useful, though imperfect, confidence signal. Similar trends are observed on the MMMUPro dataset across all other evaluated models (see Figure 2), supporting the hypothesis that higher consistency, both within a path and across paths for a question, is associated with an increased likelihood of correctness.

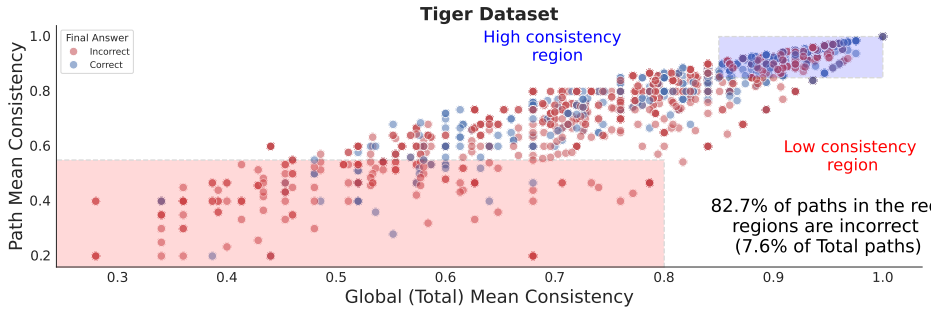

Figure 4: Scatter plot of path mean consistency versus global (total) mean consistency, colored by final answer correctness.

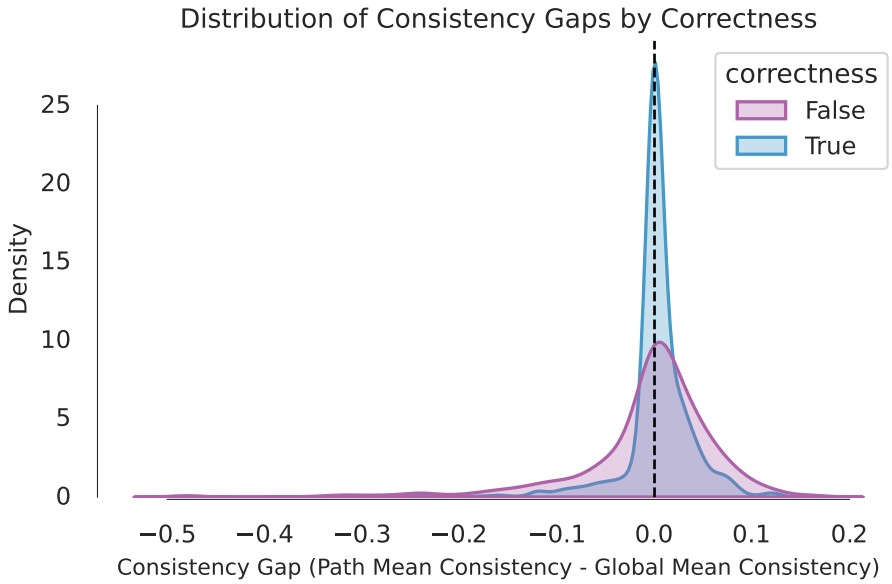

Figure 5: Distribution of consistency gap by correctness. Correct paths are slightly skewed to the right of zero, incorrect paths to the left.

Figure 5 shows the distribution of consistency gaps, defined as the difference between path-level and question-level mean consistency for TIGER-Lab/ViRL39K dataset using `Llama-4-Maverick`. Correct paths are concentrated around slightly positive gaps, indicating higher-than-average consistency, whereas incorrect paths are more dispersed and often fall below the question mean. This suggests that relative path consistency provides informative cues about the correctness of the final answer.

Figure 5 shows the distribution of consistency gaps, defined as the difference between path-level and question-level mean consistency. Positive gaps indicate that a path is more internally consistent than the average path for the same question, whereas negative gaps indicate lower relative consistency. Correct paths are concentrated around slightly positive gaps, suggesting that they tend to exhibit above-average stability across sub-questions compared to other paths for the same question. Incorrect paths, in contrast, are more widely dispersed and frequently fall below the question mean, reflecting less reliable reasoning trajectories. Overall, these results highlight systematic differences in relative consistency between correct and incorrect reasoning paths, providing a descriptive view of path-level behavior in the context of each question.

A.4 Example of First Failure Step (FFS) - First Case

We present an example of a **First Failure Step (FFS)** case, where the model's reasoning path diverges from the majority of correct paths at a specific sub-question. This example illustrates how FFS can pinpoint the earliest failure in a reasoning trajectory, enabling fine-grained analysis of error propagation. Final answer is incorrect and the first incorrect sub-question index is Q5.

Main Question:

As shown in the diagram, from point C on circle O, draw a tangent to circle O, intersecting the extension of the diameter AB at point D. If angle D equals 40 degrees, what is the measure of angle A?

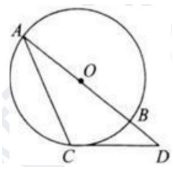

Figure 6: Diagram associated with the main question.

Sub-question + Main Question	**Path Answer / Majority Answer**
Q1: What is the measure of $\angle D$?	40 / **40**
Q2: Is CD a tangent to circle O?	Yes / **Yes**
Q3: What is the measure of $\angle OCD$?	90 / **90**
Q4: What is the measure of $\angle COD$?	50 / **50**
Q5: What is the measure of $\angle AOC$?	130 / **80** *(First Incorrect)*
Main Question: What is the measure of $\angle A$?	65 / **25**

First Failure Step (FFS): FFS identifies the earliest sub-question in a reasoning path where the model's answer diverges from the majority among correct paths, for paths that ultimately yield an incorrect final answer. In this example, the model's answer to Q5 ($\angle AOC = 130°$) deviates from the correct majority answer ($80°$), marking the first point of failure in the reasoning chain. This allows us to isolate the exact step where the model's reasoning breaks down, providing actionable insights for debugging and improving multi-step reasoning systems.

Why Knowing $\angle AOC$ Helps in Finding $\angle A$

In this example, the sub-question *"What is the measure of $\angle AOC$?"* is critical for answering *"What is the measure of $\angle A$?"* due to the **Inscribed Angle Theorem**.

Geometric Relationship:

- $\angle AOC$ is a **central angle** that subtends arc $\overset{\frown}{AC}$.
- $\angle A$ is an **inscribed angle** that subtends the same arc.
- By the Inscribed Angle Theorem:

$$\angle A = \frac{1}{2}\angle AOC$$

Example: If $\angle AOC = 80°$, then:

$$\angle A = \frac{1}{2} \times 80° = 40°$$

ARS Dependency Justification: This justifies a dependency in the ARS structure:

```
"Main Question": {
  "question": "What is the measure of angle A?",
  "depends_on_sub_question": ["Q5"],
  ...
}
```

This illustrates how sub-question dependencies reflect valid geometric reasoning steps.

A.5 Example of First Failure Step (FFS) – Second Case

We present another **First Failure Step (FFS)** case, where the model deviates from the majority reasoning at an early sub-question. In this example, the final answer is incorrect, and the first incorrect sub-question index is Q2.

Main Question:

In the Cartesian coordinate system xOy, a circle centered at the origin O passes through the point $A(13, 0)$. The line $y = kx - 3k + 4$ intersects the circle $\odot O$ at points B and C. The minimum length of the chord BC is ___.

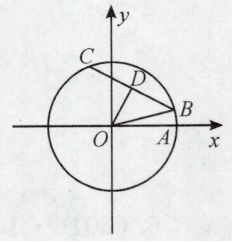

Figure 7: Diagram associated with the second FFS case.

Sub-question + Main Question	Path Answer / Majority Answer
Q1: What is the radius of the circle centered at O?	13 / **13**
Q2: What is the distance from O to the line $y = kx - 3k + 4$?	4 / **5** *(First Incorrect)*
Main Question: What is the minimum length of chord BC?	$6\sqrt{17}$ / **24**

First Failure Step (FFS): Here, the earliest deviation occurs at Q2, where the model incorrectly computed the perpendicular distance from the origin to the line (4 instead of 5). This error propagated forward, affecting both Q4 and the final answer. By identifying Q2 as the FFS, we can see how a local geometric miscalculation cascaded into a globally incorrect solution.

Why the Distance to the Line Matters

The correct computation of the distance from O to the line $y = kx - 3k + 4$ (Q2) is crucial, since it directly determines the length of chord BC.

Geometric Relationship:

- The perpendicular distance d from O to the line influences the chord length.
- If the circle radius is r, then the chord length is given by:

$$|BC| = 2\sqrt{r^2 - d^2}$$

- Using $r = 13$ and $d = 5$, the minimum chord length becomes:

$$|BC| = 2\sqrt{13^2 - 5^2} = 2\sqrt{169 - 25} = 2\sqrt{144} = 24$$

ARS Dependency Justification: This shows why Q2 is critical for answering the main question, as an incorrect distance estimate directly leads to the wrong chord length.

A.6 Example of First Failure Step (FFS) – Third Case

We present another **First Failure Step (FFS)** case, where the model's reasoning deviates at a geometric ratio step. The final answer is incorrect, and the first incorrect sub-question index is Q3.

Main Question:

As shown in the figure, in $\triangle ABC$, E is the midpoint of AC, and point D is a point on BC such that $BD : CD = 2 : 3$. Lines AD and BE intersect at point O. If $S_{\triangle AOE} - S_{\triangle BOD} = 1$, then the area of $\triangle ABC$ is ___.

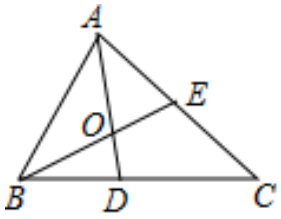

Figure 8: Diagram of $\triangle ABC$ with midpoints and ratios.

Sub-question + Main Question	**Path Answer / Majority Answer**
Q1: What is the ratio of BD to CD?	$2/3$ / $2/3$
Q2: Is E the midpoint of AC?	Yes / **Yes**
Q3: What is the ratio of the areas of $\triangle AOE$ and $\triangle BOD$?	4 / $5/4$ *(First Incorrect)*
Q6: What is the area of $\triangle ABD$?	$56/9$ / **4**
Q5: What is the area of $\triangle ABE$?	6 / **5**
Main Question: What is the area of $\triangle ABC$?	$140/9$ / **10**

First Failure Step (FFS): The earliest failure occurs at Q3, where the model incorrectly computed the ratio of areas between $\triangle AOE$ and $\triangle BOD$. The correct majority answer is $5/4$, but the model gave 4. This miscalculation distorted subsequent area computations, ultimately leading to an incorrect total area for $\triangle ABC$.

Why the Ratio of Areas Determines the Final Answer

The ratio $\dfrac{S_{\triangle AOE}}{S_{\triangle BOD}}$ is a key step in linking the given condition

$$S_{\triangle AOE} - S_{\triangle BOD} = 1$$

to the absolute area of $\triangle ABC$.

Geometric Relationship:

- The intersection of cevians AD and BE defines proportional sub-triangles.
- Correct computation of the ratio $\frac{5}{4}$ allows us to solve for the absolute areas of both $\triangle AOE$ and $\triangle BOD$.
- From these, one can scale up to the entire $\triangle ABC$, yielding:

$$S_{\triangle ABC} = 10$$

ARS Dependency Justification: Since the main condition involves the difference of two sub-triangle areas, Q3 directly governs the chain of reasoning. An incorrect ratio propagates to all later computations.

A.7 Example of First Failure Step (FFS) – Fourth Case

We present another **First Failure Step (FFS)** case, where the model's reasoning diverges at the coordinate computation of a key point. The final answer is incorrect, and the first incorrect sub-question index is Q5.

Main Question:

As shown in the figure, the sides OA and OC of rectangle $OABC$ lie on the x-axis and y-axis, respectively, and the coordinates of point B are $(3, 2)$. Points D and E are on sides AB and BC, respectively, with $BD = BE = 1$. When $\triangle BDE$ is folded along line DE, point B lands at point B'. What are the coordinates of B'?

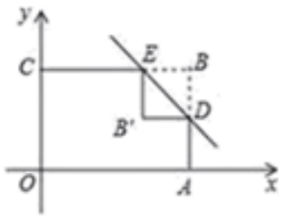

Figure 9: Diagram of rectangle $OABC$ and folding construction.

Sub-question + Main Question	**Path Answer / Majority Answer**
Q1: What are the coordinates of point B?	$(3, 2)$ / $(3, 2)$
Q2: What is the length of BD?	1 / 1
Q3: What is the length of BE?	1 / 1
Q4: What are the coordinates of point D?	$(3, 1)$ / $(3, 1)$
Q5: What are the coordinates of point E?	$(3, 1)$ / $(2, 2)$ *(First Incorrect)*
Q6: What is the slope of line DE?	Undefined / -1
Q7: What is the equation of line DE?	$x = 3$ / $y = -x + 4$
Main Question: What are the coordinates of point B'?	$(3, 0)$ / $(2, 1)$

First Failure Step (FFS): The earliest failure occurs at Q5, where the model incorrectly placed point E at $(3, 1)$ instead of $(2, 2)$. This mistake propagated to Q6 and Q7 (incorrect slope and line equation), ultimately yielding the wrong folded position of B'. Identifying Q5 as the FFS shows how a single misstep in coordinate placement cascades into a completely wrong construction.

Why Correct Placement of E is Critical

Point E lies on side BC, one unit away from $B(3, 2)$. Correct placement gives $E(2, 2)$, not $(3, 1)$.

Key Consequences:

- With $E(2, 2)$ and $D(3, 1)$, line DE has slope -1 and equation $y = -x + 4$.
- Folding across DE maps $B(3, 2)$ to its reflection $B'(2, 1)$.
- Thus, the correct answer is:

$$B' = (2, 1)$$

ARS Dependency Justification: Since E defines line DE, an incorrect coordinate at Q5 propagates to the slope, line equation, and ultimately the reflection point B'. This demonstrates how a local coordinate error derails the entire reasoning path.