



Math LLM Evaluation Suite

Single Question Mode Dataset Evaluation Mode Visualize Auto-Loop Results

Dataset Evaluation — Visual Summary

Raw Results Table

ID	Raw LLM Answer	Python Output	Dataset Answer	Dataset Comparison	Python Correct?	LLM v:
33	test/algebra/1072.json	To find the eighth term	243/625	$\frac{243}{625}$	Match	<input checked="" type="checkbox"/>
34	test/counting_and_probability/119.json	The constant term in th	nan	-125	Mismatch + explanation	<input type="checkbox"/>
35	test/number_theory/627.json	To solve the problem, w	[]	3	Match	<input checked="" type="checkbox"/>
36	test/intermediate_algebra/428.json	The roots of the equati	[3, 5, 7]	3, 5, 7	Match	<input checked="" type="checkbox"/>
37	test/geometry/967.json	72	72.0	72	Match	<input checked="" type="checkbox"/>
38	test/algebra/24.json	2000 calories.	2000.000000000000	2000	Match	<input checked="" type="checkbox"/>
39	test/number_theory/45.json	55	23	23	Mismatch + explanation	<input type="checkbox"/>
40	test/prealgebra/930.json	12 cm	12	12	Match	<input checked="" type="checkbox"/>
41	test/geometry/627.json	18	5.000000000000000	17	Mismatch + explanation	<input type="checkbox"/>
42	test/algebra/2214.json	To solve the compound		4	Error	<input type="checkbox"/>
43	test/number_theory/145.json	50	701.145767110038	701.145767110038	Error	<input type="checkbox"/>

Summary Metrics

Total Questions

100

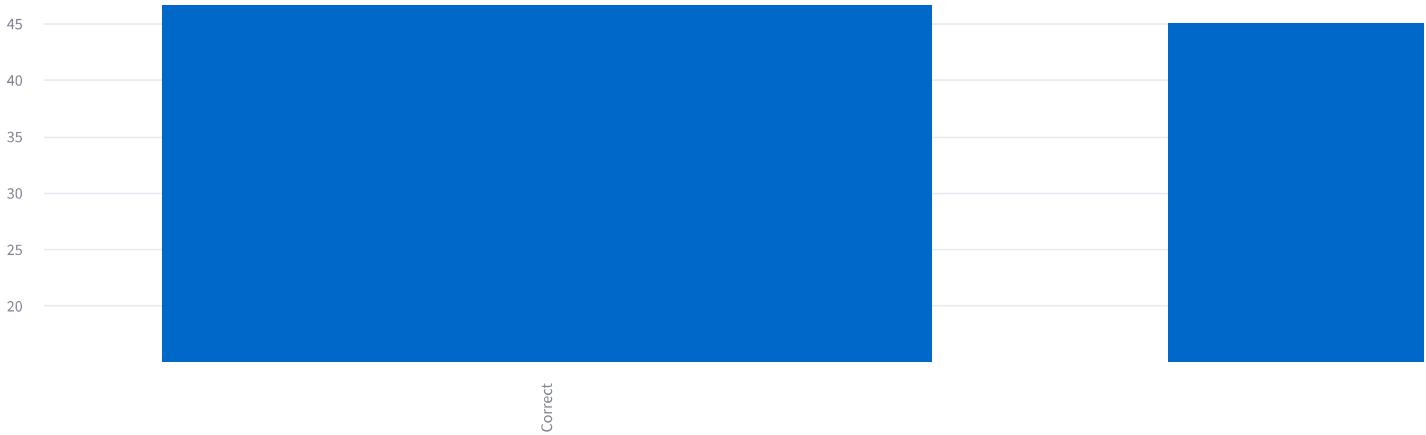
Correct Python Outputs

55

Python Accuracy (%)

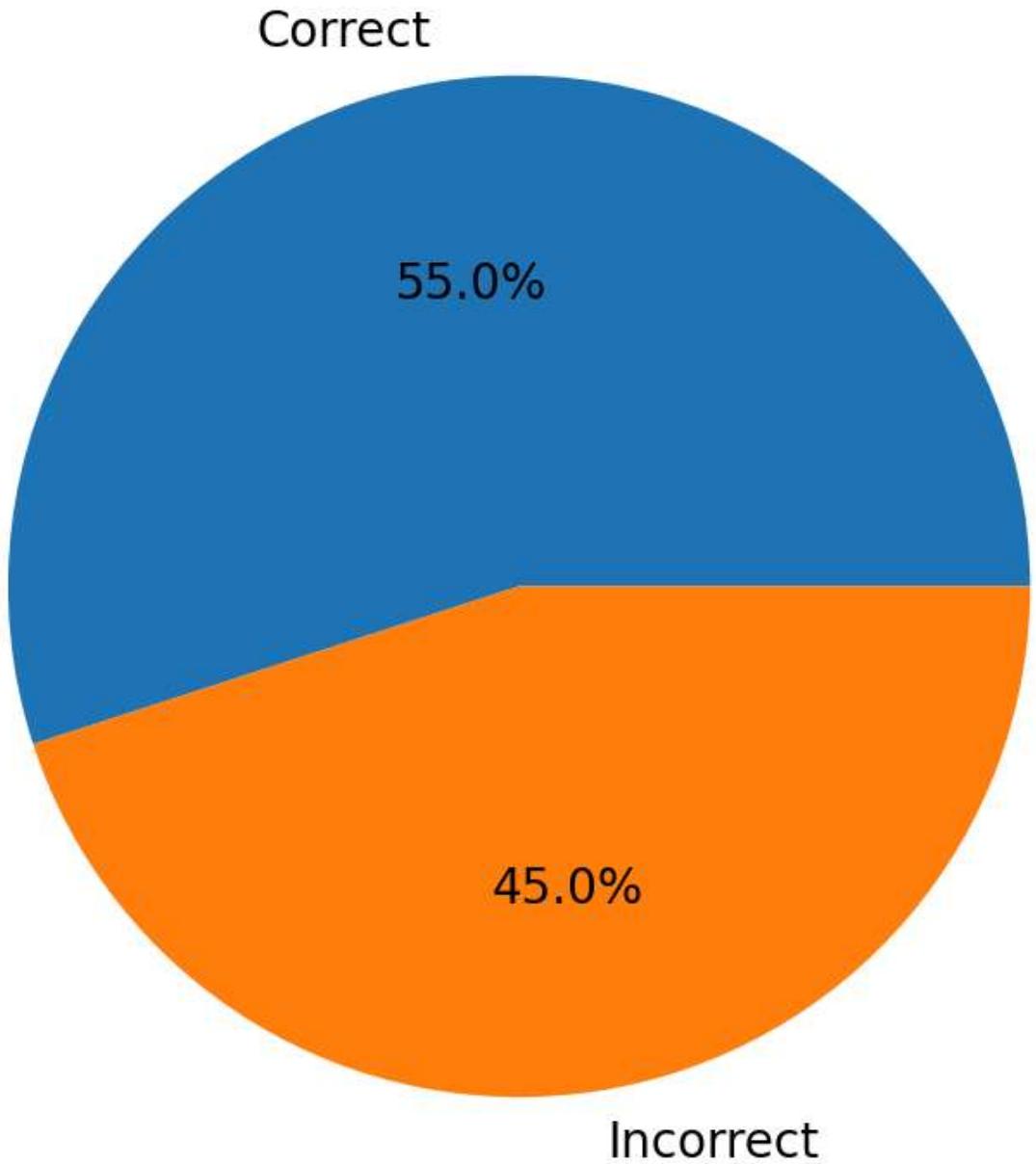
55.00%

Correct vs Incorrect Predictions

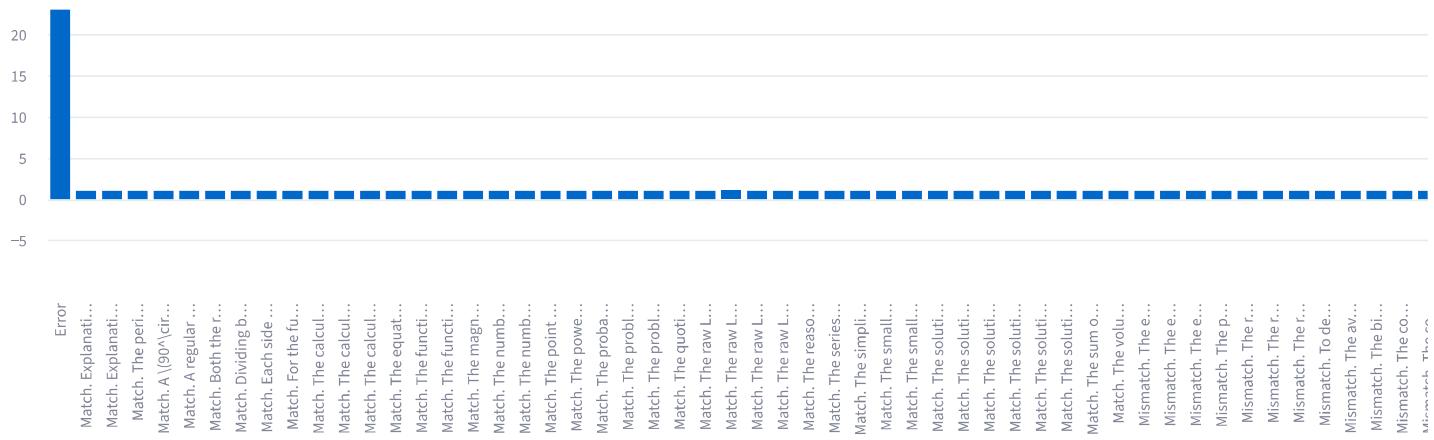


Distribution

Correctness Distribution



LLM vs Python Match Distribution



Filter Incorrect Predictions

45 incorrect predictions found:

		Raw LLM Answer	Python Output	Dataset Answer	Dataset Comparison
85	etry/353.json	The radius of the tank is $\sqrt{5}$ meters.	$\sqrt{5}$	Error	
87	a/567.json	The largest value of x that satisfies the equation is $x = 2$.	1	Error	
89	er_theory/357.json	3	21	Error	
90	ebara/1761.json	15	15.0	$\frac{3}{2}$	Mismatch + explanation
91	a/2023.json	3	3	1	Mismatch + explanation
92	ng_and_probability/377.json	The probability that exactly 4 of the islands have treasure is given by the binomial probability formula: $P(X=4) = \binom{10}{4} \cdot p^4 \cdot (1-p)^6$. Substituting $p = 0.05$, we get $P(X=4) = \binom{10}{4} \cdot 0.05^4 \cdot 0.95^6 \approx 0.02867200000000000$.	0.02867200000000000	$\frac{448}{15625}$	Mismatch + explanation
94	ebara/1646.json	The degree measure of angle AFD is 70°.	70.0	80	Mismatch + explanation
95	culus/34.json	$y = 1$	-4	Error	
96	mediate_algebra/662.json	-1, 3	[-3.35889894354067, 5]	$\pm \sqrt{19}$	Mismatch + explanation
99	culus/1300.json	$\begin{pmatrix} v \\ 1 & 2 & 3 \end{pmatrix}$	$\begin{pmatrix} v \\ -1 & 3 \end{pmatrix}$	$\begin{pmatrix} v \\ 1 & 2 & 3 \end{pmatrix}$	Error