



Math LLM Evaluation Suite

🚩 Single Question Mode 📁 Dataset Evaluation Mode 📊 Visualize Auto-Loop Results



Dataset Evaluation — Visual Summary

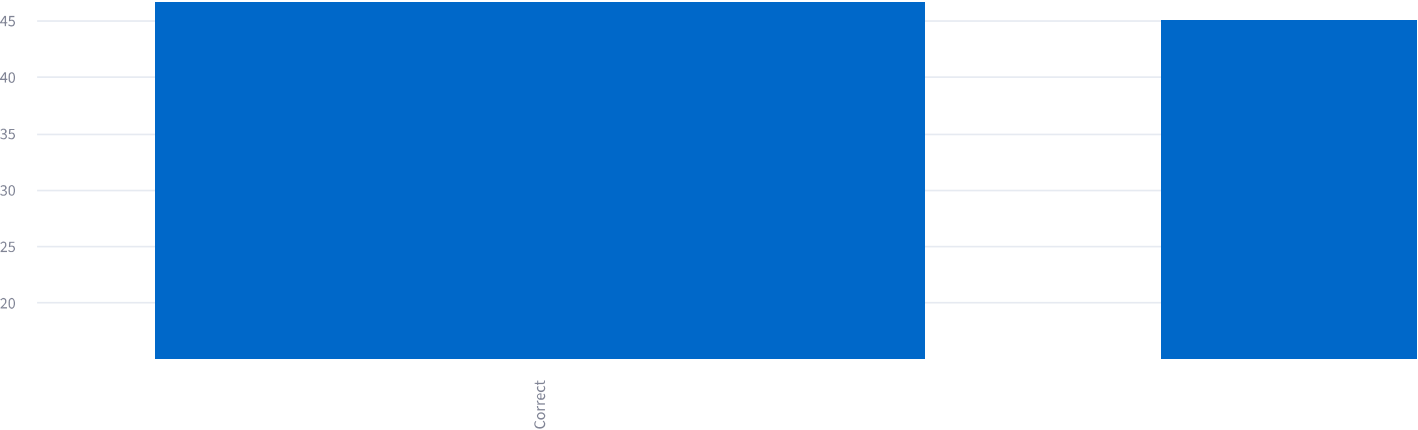
Raw Results Table

	ID	Raw LLM Answer	Python Output	Dataset Answer	Dataset Comparison	Python Correct?	LLM vs Dataset
33	test/algebra/1072.json	To find the eighth term	243/625	$\frac{243}{625}$	Match	<input checked="" type="checkbox"/>	Mismatch
34	test/counting_and_probability/119.json	The constant term in the	nan	-125	Mismatch + explanation	<input type="checkbox"/>	Mismatch
35	test/number_theory/627.json	To solve the problem, we	[]	3	Match	<input checked="" type="checkbox"/>	Match
36	test/intermediate_algebra/428.json	The roots of the equation	[3, 5, 7]	3, 5, 7	Match	<input checked="" type="checkbox"/>	Mismatch
37	test/geometry/967.json	72	72.0	72	Match	<input checked="" type="checkbox"/>	Match
38	test/algebra/24.json	2000 calories.	2000.0000000000000	2000	Match	<input checked="" type="checkbox"/>	Match
39	test/number_theory/45.json	55	23	23	Mismatch + explanation	<input type="checkbox"/>	Mismatch
40	test/prealgebra/930.json	12 cm	12	12	Match	<input checked="" type="checkbox"/>	Match
41	test/geometry/627.json	18	5.0000000000000000	17	Mismatch + explanation	<input type="checkbox"/>	Mismatch
42	test/algebra/2214.json	To solve the compound		4	Error	<input type="checkbox"/>	Error
43	test/intermediate_algebra/1454.json	50		$70\sqrt{2}+92$	Error	<input type="checkbox"/>	Error

Summary Metrics

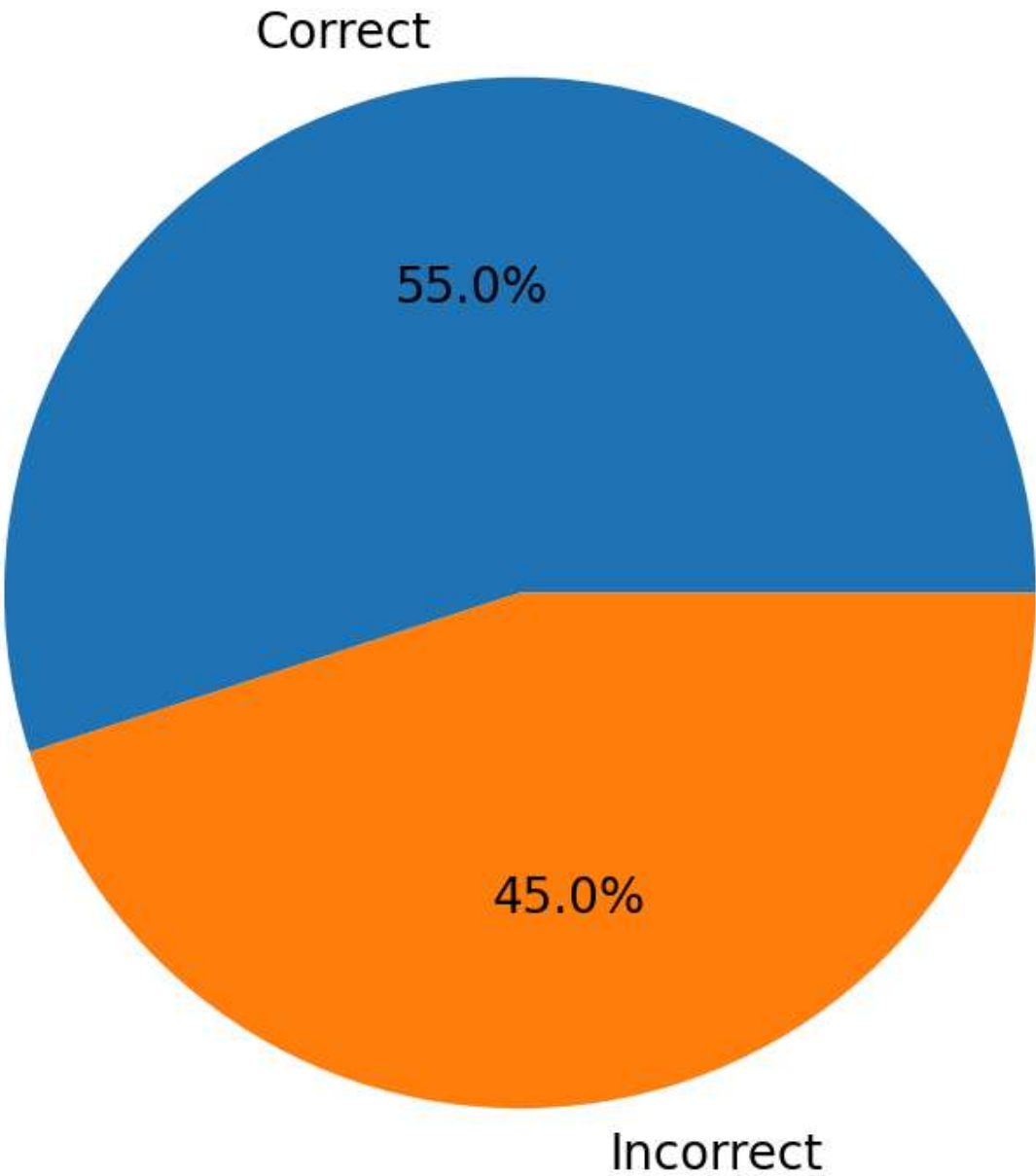
Total Questions	Correct Python Outputs	Python Accuracy (%)
100	55	55.00%

Correct vs Incorrect Predictions

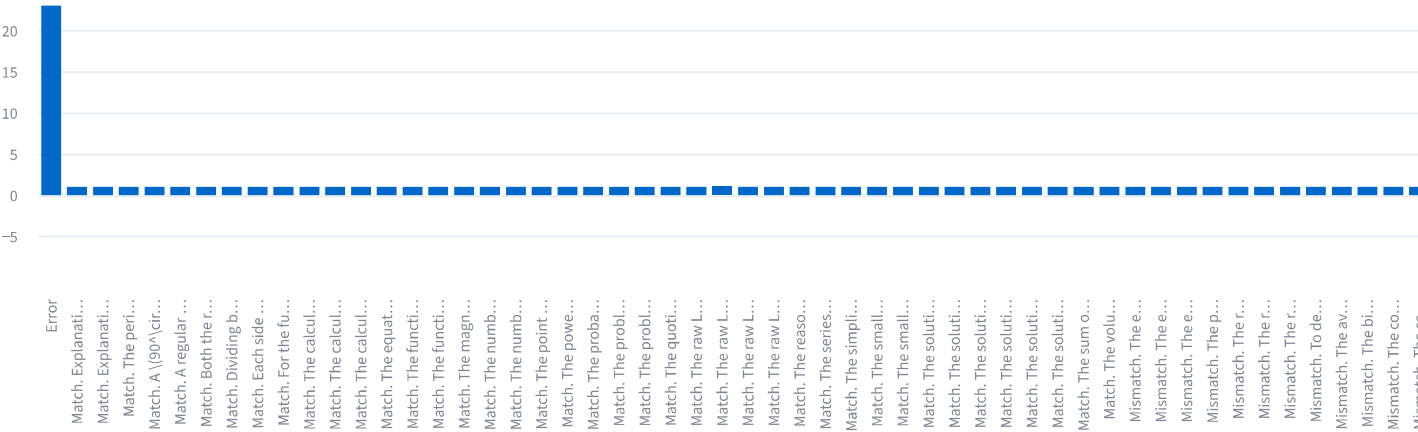


Distribution

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	ebra/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: $\binom{10}{4} (0.2)^4 (0.8)^6$.	0.028672000000000	$\frac{448}{15625}$	Mismatch + explanation
94	ebra/1646.json	The degree measure of angle $\angle AFD$ is 70° .	70.0	80	Mismatch + explanation
95	culus/34.json	$y = 1$		-4	Error
96	ediate_algebra/662.json	-1, 3	$[-3.35889894354067, 5.1]$	$\pm \sqrt{19}$	Mismatch + explanation
99	culus/1300.json	$\mathbf{v} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$		$\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$	Error