



Pass@1 Range	Total Problems	Number of CoT graded	Number of problems graded	Correct CoT	Incorrect CoT
0-10%	5	12	4	11	1
10-25%	10	30	10	27	3
25-75%	29	30	10	29	1
75-100%	84	30	10	30	0

Grade	Description				
pass	The chain of thought reasoning is correct. There are no mistakes, and the answer is correctly derived.				
fail	There are mistakes in the chain of thought				
incorrect ground truth	The ground truth answer is incorrect.				