	GSM8K	Num. Sub.	Digit Exp.	IDF Conv.	Add. Op.	Rev. Op.	Prob. Underst.	Crit. Thinking	Dist. Ins.
Human	96.8	92.9 (4.0)	100.0 (-3.3)	100.0 (-3.3)	87.5 (9.6)	100.0 (-3.3)	100.0 (-3.3)	100.0 (-3.3)	100.0 (-3.3)
GPT-4	93.2	89.8 (3.7)	90.5 (3.0)	89.0 (4.5)	79.5 (14.7)	83.7 (10.2)	93.9 (-0.7)	90.8 (2.7)	67.5 (27.6)
GPT-3.5-Turbo	73.6	69.5 (5.6)	70.4 (4.4)	62.3 (15.3)	48.5 (34.2)	55.2 (25.0)	74.2 (-0.8)	62.2 (15.6)	47.3 (35.7)
Mistral-7B	39.6	35.2 (11.1)	35.9 (9.2)	29.9 (24.5)	14.4 (63.6)	21.8 (45.0)	38.7 (2.3)	28.1 (29.1)	5.5 (86.0)
LLaMA-2-7B	13.4	13.0 (3.4)	10.0 (25.4)	10.4 (22.6)	3.0 (77.4)	7.0 (47.5)	13.9 (-4.0)	7.6 (43.5)	0.0 (100.0)
CodeLlama-7B	25.3	22.3 (12.0)	23.8 (6.0)	19.1 (24.5)	8.6 (66.2)	9.3 (63.1)	25.9 (-2.4)	11.4 (55.1)	1.4 (94.3)
LLaMA-2-13B	25.4	25.9 (-2.1)	22.9 (9.8)	17.7 (30.2)	9.5 (62.7)	13.4 (47.2)	27.4 (-7.8)	15.4 (39.4)	0.3 (98.8)
CodeLlama-13B	35.9	29.6 (17.7)	29.9 (16.7)	28.2 (21.5)	14.6 (59.5)	15.4 (57.2)	34.6 (3.6)	16.7 (53.6)	4.4 (87.8)
CodeLlama-34B	45.6	29.6 (35.2)	29.9 (34.4)	28.2 (38.2)	14.6 (68.1)	15.4 (66.3)	34.6 (24.1)	16.7 (63.5)	4.4 (90.4)
LLaMA-2-70B	56.7	53.3 (6.0)	53.1 (6.4)	42.1 (25.8)	31.5 (44.4)	36.6 (35.4)	56.6 (0.3)	46.9 (17.4)	0.3 (99.5)
MetaMath-Mistral	77.8	71.0 (8.7)	70.0 (10.0)	61.9 (20.4)	45.1 (42.0)	58.1 (25.2)	77.5 (0.4)	55.6 (28.5)	10.6 (86.4)
MetaMath-7B	66.7	59.1 (11.4)	58.8 (11.9)	49.7 (25.6)	30.9 (53.8)	49.7 (25.6)	64.9 (2.7)	36.7 (45.0)	5.2 (92.3)
Abel-7B	59.5	56.1 (5.7)	51.0 (14.4)	38.6 (35.1)	24.6 (58.6)	33.5 (43.7)	58.7 (1.4)	33.3 (44.1)	1.0 (98.3)
ToRA-7B	67.5	62.0 (8.1)	64.8 (3.9)	54.1 (19.8)	32.2 (52.2)	41.4 (38.7)	68.2 (-1.1)	26.0 (61.5)	0.0 (100.0)
MAmmoTH-7B	52.8	45.2 (14.5)	49.3 (6.7)	38.2 (27.7)	21.5 (59.4)	26.8 (49.4)	51.8 (2.0)	24.4 (53.8)	0.0 (100.0)
MAmmoTH-Coder-7B	59.9	54.8 (8.5)	56.6 (5.4)	45.8 (23.5)	29.0 (51.5)	31.5 (47.5)	58.5 (2.4)	33.7 (43.8)	0.0 (100.0)
SEGO-7B	68.7	60.4 (12.1)	64.3 (6.4)	51.7 (24.7)	35.9 (47.8)	41.0 (40.3)	67.2 (2.1)	37.2 (45.8)	0.0 (100.0)
MetaMath-13B	70.8	61.5 (13.2)	64.3 (9.2)	53.1 (24.9)	36.3 (48.7)	53.9 (23.9)	71.7 (-1.2)	42.9 (39.4)	4.9 (93.0)
Abel-13B	66.7	62.4 (6.5)	59.7 (10.5)	50.0 (25.1)	34.8 (47.8)	41.6 (37.6)	67.3 (-0.9)	45.5 (31.8)	1.8 (97.3)
ToRA-13B	71.8	65.3 (9.0)	67.8 (5.5)	56.7 (21.0)	39.9 (44.5)	45.8 (36.2)	72.7 (-1.3)	34.7 (51.6)	0.0 (100.0)
MAmmoTH-13B	62.4	54.9 (12.0)	58.5 (6.2)	48.7 (22.0)	31.4 (49.7)	34.1 (45.3)	61.9 (0.8)	37.1 (40.6)	0.0 (100.0)
MAmmoTH-Coder-13B	64.9	59.4 (8.5)	62.0 (4.4)	50.3 (22.5)	36.9 (43.1)	36.2 (44.3)	63.8 (1.6)	43.2 (33.4)	0.0 (100.0)
SEGO-13B	72.5	65.5 (9.7)	68.5 (5.5)	58.6 (19.2)	43.6 (39.9)	45.9 (36.6)	71.6 (1.3)	40.6 (43.9)	0.0 (100.0)
MetaMath-70B	82.1	74.9 (8.8)	74.5 (9.2)	65.0 (20.9)	51.0 (37.9)	58.0 (29.4)	79.6 (3.0)	61.9 (24.6)	10.0 (87.8)
Abel-70B	83.8	76.7 (8.6)	76.9 (8.3)	63.6 (24.1)	53.1 (36.6)	60.0 (28.4)	81.4 (2.9)	64.8 (22.7)	3.0 (96.5)
MAmmoTH-70B	75.9	67.4 (11.2)	71.7 (5.6)	59.2 (22.0)	47.8 (37.1)	49.1 (35.3)	75.5 (0.5)	56.6 (25.4)	0.0 (100.0)