Model Performance and Compute Summary (Top & Bottom Models)

Model	F1	Acc	Prec	Recall	Exec (s)	Exec SD (s)	Eval Mean (s)	Prompt Count	Eval Count
Llama3.1 (8b) (system1)	0.5870	62.0000	0.4576	0.8182	0.2795	0.7230	162510000.0000	248.68	17.00
Falcon3 (7b) (system1)	0.5682	62.0000	0.4630	0.7353	0.3150	0.5265	221770000.0000	267.51	26.00
Gemma2 (9b) (cot_nshot)	0.5625	58.0000	0.4355	0.7941	4.0741	0.5169	3457570000.0000	2048.00	276.35
Athene-v2 (72b) (cot_nshot)	0.5517	63.8889	0.4706	0.6667	58.5299	7.5095	74082708333.3333	2048.00	362.76
Olmo2 (7b) (system1)	0.5479	67.0000	0.4000	0.8696	0.1993	0.3688	124340000.0000	249.70	13.00
Athene-v2 (72b) (cot)	0.5455	60.0000	0.4138	0.8000	62.1297	6.6110	61464570000.0000	330.60	425.93
Qwen2.5 (72b) (cot_nshot)	0.5432	63.0000	0.5000	0.5946	57.9709	6.4865	73533320000.0000	2048.00	359.62
Marco-o1 (7b) (cot_nshot)	0.5361	55.0000	0.3714	0.9630	2.8180	0.3715	2383310000.0000	2048.00	271.72
Athene-v2 (72b) (system1)	0.5333	65.0000	0.4082	0.7692	3.4656	4.6348	2480600000.0000	251.95	18.00
Llama3.3 (70b) (cot_nshot)	0.5278	66.0000	0.4318	0.6786	26.4581	3.3192	12981990000.0000	2048.00	380.90
Olmo2 (7b) (cot)	0.5191	37.0000	0.3505	1.0000	2.8126	0.6787	2770260000.0000	326.85	329.81
ExaOne3.5 (7.8b) (cot_nshot)	0.5185	48.0000	0.3836	0.8000	2.9127	0.3187	2437790000.0000	2048.00	277.16
Llama3.2 (3b) (system1)	0.5161	70.0000	0.5000	0.5333	0.1850	0.4088	100950000.0000	258.39	17.00
Qwen2.5 (3b) (system1)	0.5143	66.0000	0.4390	0.6207	0.2026	0.3337	114160000.0000	251.92	18.00
Command-R (35b) (cot)	0.5128	43.0000	0.3529	0.9375	9.9704	1.8151	9849100000.0000	331.94	352.51
Phi4 (14b) (cot_nshot)	0.3571	46.0000	0.2344	0.7500	8.1288	1.8470	7301570000.0000	2048.00	494.78
Qwen2.5 (0.5b) (cot)	0.3500	48.0000	0.2800	0.4667	0.5781	0.2979	529590000.0000	330.70	178.25
Llama3.2 (1b) (cot)	0.3419	23.0000	0.2062	1.0000	0.6680	0.3449	631880000.0000	336.98	232.58
Qwen2.5 (3b) (cot_nshot)	0.3243	50.0000	0.2553	0.4444	1.8726	0.3612	1562560000.0000	2048.00	258.62
Llama3.2 (1b) (system1)	0.3038	45.0000	0.2182	0.5000	0.1211	0.2114	60800000.0000	257.44	18.44
Nemotron Mini (4b) (system1)	0.2963	62.0000	0.2857	0.3077	0.2401	0.4050	139400000.0000	261.84	19.90
Mistral (7b) (cot)	0.2941	52.0000	0.2381	0.3846	2.9245	1.0996	2887590000.0000	363.84	366.01
Qwen2.5 (7b) (system1)	0.2745	63.0000	0.3333	0.2333	0.2787	0.6334	164870000.0000	252.04	18.00
Tulu3 (8b) (system1)	0.2647	50.0000	0.2000	0.3913	0.2498	0.7492	130290000.0000	249.70	13.00
Marco-o1 (7b) (system1)	0.2553	65.0000	0.3000	0.2222	0.2849	0.7164	165200000.0000	250.05	18.00
Qwen2.5 (1.5b) (cot)	0.2500	64.0000	0.3529	0.1935	1.4244	0.4420	1374350000.0000	330.73	347.66
Qwen2.5 (0.5b) (system1)	0.2000	52.0000	0.2069	0.1935	0.1385	0.1974	67360000.0000	252.61	18.41
Qwen2.5 (14b) (system1)	0.1702	61.0000	0.1538	0.1905	0.4627	1.0219	283130000.0000	251.46	18.00
Mistral (7b) (cot_nshot)	0.0984	45.0000	0.0857	0.1154	2.4661	0.7119	2026680000.0000	2048.00	236.91
Smallthinker (3b) (cot_nshot)	0.0000	0.0000	0.0000	0.0000	5.3261	0.0000	4838000000.0000	2048.00	708.00