Model Performance and Compute Summary (Top & Bottom Models)

model	f1_score	accuracy	precision	recall	exec_time_mean	eval_time_mean	eval_duration_std	prompt_eval_count_mean	eval_count_mean
llama3.1:8b-instruct-q4_K_M_(system1)	0.5870	62.0000	0.4576	0.8182	0.2795	0.7230	162510000.0000	248.68	17.00
falcon3:7b-instruct-q4_K_M_(system1)	0.5682	62.0000	0.4630	0.7353	0.3150	0.5265	221770000.0000	267.51	26.00
gemma2:9b-instruct-q4_K_M_(cot_nshot)	0.5625	58.0000	0.4355	0.7941	4.0741	0.5169	3457570000.0000	2048.00	276.35
athene-v2:72b-q4_K_M_(cot_nshot)	0.5517	63.8889	0.4706	0.6667	58.5299	7.5097	54082708333.3333	2048.00	362.76
olmo2:7b-1124-instruct-q4_K_M_(system1)	0.5479	67.0000	0.4000	0.8696	0.1993	0.3688	124340000.0000	249.70	13.00
athene-v2:72b-q4_K_M_(cot)	0.5455	60.0000	0.4138	0.8000	62.1297	6.6110	61464570000.0000	330.60	425.93
qwen2.5:72b-instruct-q4_K_M_(cot_nshot)	0.5432	63.0000	0.5000	0.5946	57.9709	6.4867	53533320000.0000	2048.00	359.62
marco-o1:7b-q4_K_M_(cot_nshot)	0.5361	55.0000	0.3714	0.9630	2.8180	0.3715	2383310000.0000	2048.00	271.72
athene-v2:72b-q4_K_M_(system1)	0.5333	65.0000	0.4082	0.7692	3.4656	4.6348	2480600000.0000	251.95	18.00
llama3.3:70b-instruct-q4_K_M_(cot_nshot)	0.5278	66.0000	0.4318	0.6786	26.4581	3.3191	22981990000.0000	2048.00	380.90
nemotron-mini:4b-instruct-q4_K_M_(system1)	0.2963	62.0000	0.2857	0.3077	0.2401	0.4050	139400000.0000	261.84	19.90
mistral:7b-instruct-q4_K_M_(cot)	0.2941	52.0000	0.2381	0.3846	2.9245	1.0996	2887590000.0000	363.84	366.01
qwen2.5:7b-instruct-q4_K_M_(system1)	0.2745	63.0000	0.3333	0.2333	0.2787	0.6334	164870000.0000	252.04	18.00
tulu3:8b-q4_K_M_(system1)	0.2647	50.0000	0.2000	0.3913	0.2498	0.7492	130290000.0000	249.70	13.00
marco-o1:7b-q4_K_M_(system1)	0.2553	65.0000	0.3000	0.2222	0.2849	0.7164	165200000.0000	250.05	18.00
qwen2.5:1.5b-instruct-q4_K_M_(cot)	0.2500	64.0000	0.3529	0.1935	1.4244	0.4420	1374350000.0000	330.73	347.66
qwen2.5:0.5b-instruct-q4_K_M_(system1)	0.2000	52.0000	0.2069	0.1935	0.1385	0.1974	67360000.0000	252.61	18.41
qwen2.5:14b-instruct-q4_K_M_(system1)	0.1702	61.0000	0.1538	0.1905	0.4627	1.0219	283130000.0000	251.46	18.00
mistral:7b-instruct-q4_K_M_(cot_nshot)	0.0984	45.0000	0.0857	0.1154	2.4661	0.7119	2026680000.0000	2048.00	236.91
smallthinker:3b-preview-q4_K_M_(cot_nshot)	0.0000	0.0000	0.0000	0.0000	5.3261	0.0000	4838000000.0000	2048.00	708.00