

Open LLM Leaderboard v1/v2 (Average error across benchmarks)

FLOPs	6.7	7.3	19.7	9.2	17.2	6.1	11.6
Size and Tokens	2.1	3.9	7.2	7.9	8.2	3.0	4.3
PCA + FLOPs (d=1)	3.0	6.7	4.0	5.5	26.6	5.5	10.5
PCA + FLOPs (d=2)	3.4	5.1	3.4	5.2	25.4	5.9	9.8
PCA + FLOPs (d=3)	3.7	5.9	3.1	5.5	24.9	6.0	9.9
PCA + FLOPs (d=4)	3.6	5.3	4.0	6.8	24.8	5.9	10.0
Sloth basic (d=1) (shared intercept)	3.0	7.8	4.2	5.6	6.0	10.3	6.1
Sloth basic (d=2) (shared intercept)	2.5	6.6	2.3	5.7	6.5	10.9	6.5
Sloth basic (d=3) (shared intercept)	2.5	6.6	2.6	5.5	8.1	10.8	6.2
Sloth basic (d=4) (shared intercept)	2.5	6.5	2.5	5.6	7.7	10.9	6.1
Sloth (d=1) (shared intercept)	2.8	7.0	4.7	6.6	6.1	10.6	6.6
Sloth (d=2) (shared intercept)	2.3	6.1	2.7	6.2	5.1	13.0	6.3
Sloth (d=3) (shared intercept)	3.0	5.4	3.6	6.3	5.3	12.8	6.4
Sloth (d=4) (shared intercept)	2.7	5.4	3.4	6.2	4.3	11.3	6.4
Sloth basic (d=1)	3.6	6.8	4.3	10.5	9.8	6.6	5.4
Sloth basic (d=2)	3.5	6.2	4.7	8.7	7.0	4.7	5.1
Sloth basic (d=3)	3.0	5.4	5.2	9.0	5.9	4.3	4.8
Sloth basic (d=4)	2.1	4.9	2.4	9.3	7.1	4.6	4.8
Sloth (d=1)	3.7	7.7	6.2	6.7	8.4	7.5	3.8
Sloth (d=2)	3.3	6.9	4.4	6.6	4.0	4.2	4.2
Sloth (d=3)	2.2	3.9	3.9	7.5	4.8	2.7	3.1
Sloth (d=4)	1.6	4.4	3.4	6.8	4.8	4.1	2.5
	bloom	llama-2	pythia	qwen1.5	qwen2	starcoder2	yi-1.5