Open LLM Leaderboard v1 (Average error across benchmarks)

| | bloom | codegen-nl | codellama | deepseek-coder-base | pythia | falcon | gemma | gpt-j-neo-neox | internlm2 | meta-llama-3 | mpt | olmo | opt | qwen2 | rwkv-4-pile | starcoder2 | stablelm-base-alpha | xglm | yi-1.5 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| FLOPs (shared intercept) | 5.3 | 2.6 | 12.2 | 13.1 | 2.6 | 9.1 | 8.0 | 2.4 | 12.3 | 10.5 | 4.0 | 7.3 | 3.9 | 9.8 | 2.4 | 13.8 | 12.7 | 5.1 | 11.0 |
| FLOPs | 4.4 | 2.9 | 6.4 | 6.9 | 5.1 | 9.1 | 9.1 | 7.8 | 3.9 | **2.9** | 4.6 | **0.8** | 6.9 | 8.7 | 4.0 | **3.2** | 1.3 | 3.0 | 2.3 |
| Size and Tokens | 2.8 | 2.8 | **5.3** | 6.7 | 7.7 | 6.2 | 6.9 | 5.6 | **2.4** | 3.3 | 4.9 | 4.6 | 6.5 | 5.2 | **1.3** | 6.9 | 1.2 | 2.0 | 1.8 |
| PCA + FLOPs (d=1) | 5.3 | 6.6 | 8.0 | 4.7 | 7.9 | 10.2 | 10.8 | 8.8 | 5.3 | 6.3 | 8.5 | 7.9 | 9.4 | 12.7 | 6.6 | 7.3 | 2.7 | 4.8 | 5.3 |
| PCA + FLOPs (d=2) | 5.4 | 2.2 | 7.1 | **3.1** | 6.7 | 9.2 | 10.2 | 8.5 | 4.4 | 6.2 | 3.6 | 4.1 | 9.4 | 13.6 | 6.4 | 4.5 | 2.0 | 4.7 | 3.5 |
| PCA + FLOPs (d=3) | 5.5 | **1.7** | 7.0 | 3.4 | 7.5 | 11.2 | 10.5 | 9.5 | 3.4 | 6.4 | 4.7 | 2.9 | 10.3 | 13.9 | 6.9 | 4.5 | 2.0 | 5.3 | 3.9 |
| PCA + FLOPs (d=4) | 5.6 | 2.3 | 7.0 | 3.9 | 7.6 | 11.1 | 10.6 | 9.7 | 3.0 | 5.8 | 4.6 | 2.9 | 10.4 | 14.4 | 7.0 | 4.2 | 2.1 | 5.3 | 3.8 |
| Sloth basic (d=1) (shared intercept) | 6.6 | 2.8 | 12.5 | 13.0 | 3.0 | 8.6 | 6.4 | 2.6 | 12.2 | 9.6 | 4.1 | 8.1 | 4.0 | 8.7 | 2.6 | 14.0 | 12.5 | 4.5 | 10.5 |
| Sloth basic (d=2) (shared intercept) | 4.5 | **1.7** | 11.9 | 12.1 | 2.2 | 8.1 | 6.9 | 2.3 | 12.1 | 9.8 | 3.2 | 9.2 | 3.5 | 5.6 | 2.1 | 13.0 | 12.9 | 4.9 | 9.4 |
| Sloth basic (d=3) (shared intercept) | 4.9 | **1.7** | 13.0 | 12.5 | 1.8 | 8.6 | 6.8 | 2.3 | 11.9 | 13.6 | 2.7 | 9.2 | 3.2 | 5.3 | 2.1 | 13.1 | 12.8 | 4.9 | 9.2 |
| Sloth basic (d=4) (shared intercept) | 5.0 | **1.7** | 13.0 | 12.5 | 1.8 | 8.6 | 6.8 | 2.3 | 11.9 | 13.6 | 2.6 | 9.2 | 3.2 | 5.3 | 2.1 | 13.1 | 12.8 | 4.9 | 9.2 |
| Sloth (d=1) (shared intercept) | 2.5 | 4.6 | 15.3 | 15.1 | 2.3 | 9.8 | **5.7** | 3.0 | 6.4 | 7.4 | 5.1 | 10.9 | 4.8 | 9.7 | 2.3 | 14.7 | 14.2 | 5.6 | 6.6 |
| Sloth (d=2) (shared intercept) | 3.6 | 3.1 | 14.8 | 14.7 | 1.8 | 6.2 | 5.9 | 2.4 | 10.5 | 5.3 | 2.2 | 9.1 | 2.3 | 6.9 | 1.7 | 16.4 | 15.5 | 4.8 | 5.7 |
| Sloth (d=3) (shared intercept) | 3.1 | 2.6 | 14.4 | 14.9 | **1.5** | 6.5 | 6.3 | 2.3 | 11.7 | 5.8 | **2.1** | 10.0 | 2.2 | 6.9 | 1.5 | 17.0 | 14.7 | 4.7 | 6.1 |
| Sloth (d=4) (shared intercept) | 3.3 | 2.7 | 14.9 | 14.9 | **1.5** | 6.7 | 6.1 | 2.1 | 11.7 | 5.3 | 2.5 | 10.4 | 2.2 | 5.5 | 1.7 | 16.8 | 15.3 | 4.9 | 5.6 |
| Sloth basic (d=1) | 5.8 | 3.1 | 10.4 | 12.3 | 7.2 | **5.4** | 12.6 | 8.5 | 8.3 | 5.6 | 3.9 | 3.9 | 5.9 | 32.9 | 6.1 | 14.9 | 3.5 | 4.6 | 4.1 |
| Sloth basic (d=2) | 3.5 | 3.2 | 8.3 | 6.5 | 7.3 | 13.2 | 7.4 | 7.6 | 4.4 | 5.2 | 6.7 | 2.0 | 5.7 | 10.1 | 6.1 | 5.1 | 1.6 | 3.6 | 3.3 |
| Sloth basic (d=3) | 3.2 | 2.6 | 6.5 | 9.1 | 6.0 | 10.2 | 8.9 | 6.0 | 5.0 | 6.8 | 5.9 | 1.9 | 3.0 | 10.4 | 4.5 | 4.0 | 1.5 | 2.9 | 3.0 |
| Sloth basic (d=4) | 4.4 | 2.3 | 6.8 | 6.6 | 6.1 | 9.0 | 8.9 | 6.6 | 4.8 | 5.0 | 5.0 | 1.9 | 3.4 | 9.9 | 5.1 | 3.9 | 2.0 | 3.0 | 2.7 |
| Sloth (d=1) | 3.2 | 4.2 | 8.8 | 9.6 | 7.2 | 6.4 | 6.9 | 2.1 | 3.9 | 4.8 | 3.6 | 10.5 | 2.5 | 12.5 | 2.3 | 11.8 | 1.1 | 2.0 | 2.9 |
| Sloth (d=2) | 2.4 | 4.2 | 5.8 | 10.2 | 7.7 | 8.3 | 6.6 | 1.9 | 3.0 | 3.1 | 6.5 | 2.7 | **2.0** | **3.7** | 1.9 | 5.5 | 1.6 | 2.2 | 2.2 |
| Sloth (d=3) | 2.8 | 2.1 | **5.3** | 9.5 | 7.6 | 6.8 | 6.6 | 5.4 | 3.6 | 4.6 | 4.0 | 2.6 | **2.0** | 4.7 | **1.3** | 7.3 | 1.2 | 1.8 | **1.7** |
| Sloth (d=4) | **2.2** | 3.5 | 6.0 | 10.2 | 7.4 | 6.5 | 6.1 | **1.3** | 3.5 | 3.7 | 5.1 | 2.6 | **2.0** | 5.8 | **1.3** | 5.1 | **0.9** | **1.7** | 2.1 |