```
llama_init_from_model: KV self size  =  392.00 MiB, K (q8_0):  136.00 MiB, V (f16):  256.00 MiB
llama_init_from_model:        CPU  output buffer size =     0.49 MiB
llama_init_from_model:        CPU compute buffer size =   296.01 MiB
llama_init_from_model: graph nodes  = 1030
llama_init_from_model: graph splits = 514 (with bs=512), 1 (with bs=1)
common_init_from_params: setting dry_penalty_last_n to ctx_size = 4096
common_init_from_params: warming up the model with an empty run - please wait ... (--no-warmup to disa
ble)
main: llama threadpool init, n_threads = 32

system_info: n_threads = 32 (n_threads_batch = 32) / 64 | CPU : LLAMAFILE = 1 | AARCH64_REPACK = 1 |

sampler seed: 3873949425
sampler params:
        repeat_last_n = 64, repeat_penalty = 1.000, frequency_penalty = 0.000, presence_penalty = 0.00
0
        dry_multiplier = 0.000, dry_base = 1.750, dry_allowed_length = 2, dry_penalty_last_n = 4096
        top_k = 40, top_p = 0.950, min_p = 0.050, xtc_probability = 0.000, xtc_threshold = 0.100, typi
cal_p = 1.000, top_n_sigma = -1.000, temp = 0.800
        mirostat = 0, mirostat_lr = 0.100, mirostat_ent = 5.000
sampler chain: logits -> logit-bias -> penalties -> dry -> top-k -> typical -> top-p -> min-p -> xtc -
> temp-ext -> dist
generate: n_ctx = 4096, n_batch = 2048, n_predict = -1, n_keep = 1

Can you please count from 1 to 10?<think>
First, I'll start by listing the numbers from 1 to 10 in a clear and sequential manner.

I'll make sure each number is correctly placed and easy to read.

Finally, I'll present the final
```

```
    0[1]  4[0]  8[1] 12[1] 16[1] 20[0] 24[0] 28[0]  32[3]  36[1] 40[1] 44[0] 48[0] 52[0] 56[1] 60[1]
    1[0]  5[1]  9[1] 13[1] 17[1] 21[1] 25[0] 29[2]  33[0]  37[1] 41[5] 45[1] 49[0] 53[1] 57[0] 61[0]
    2[0]  6[0] 10[1] 14[1] 18[1] 22[0] 26[0] 30[0]  34[1]  38[1] 42[1] 46[0] 50[1] 54[0] 58[1] 62[0]
    3[1]  7[1] 11[0] 15[1] 19[0] 23[0] 27[0] 31[1]  35[1]  39[1] 43[1] 47[0] 51[0] 55[0] 59[0] 63[1]
  Mem[||||||                          1.38G/121G] Tasks: 52, 246 thr, 734 kthr; 33 running
  Swp[                                   0K/0K] Load average: 22.20 12.50 8.95
                                                Uptime: 00:17:26

  Main   I/O
    PID USER       PRI  NI  VIRT   RES   SHR S  CPU%-MEM%   TIME+  Command
   2477 lw          20   0 8377M 5161M 4691M R 100.2  4.2  1:06.11
   2574 lw          20   0 8377M 5161M 4691M R 100.2  4.2  0:54.94
   2575 lw          20   0 8377M 5161M 4691M R 100.2  4.2  0:55.01
   2576 lw          20   0 8377M 5161M 4691M R 100.2  4.2  0:55.01
   2578 lw          20   0 8377M 5161M 4691M R 100.2  4.2  0:55.01
   2579 lw          20   0 8377M 5161M 4691M R 100.2  4.2  0:55.01
   2580 lw          20   0 8377M 5161M 4691M R 100.2  4.2  0:55.01
   2584 lw          20   0 8377M 5161M 4691M R 100.2  4.2  0:54.97
   2585 lw          20   0 8377M 5161M 4691M R 100.2  4.2  0:55.00
   2586 lw          20   0 8377M 5161M 4691M R 100.2  4.2  0:55.00
   2587 lw          20   0 8377M 5161M 4691M R 100.2  4.2  0:54.98
   2588 lw          20   0 8377M 5161M 4691M R 100.2  4.2  0:55.00
   2589 lw          20   0 8377M 5161M 4691M R 100.2  4.2  0:55.01
   2590 lw          20   0 8377M 5161M 4691M R 100.2  4.2  0:55.00
   2593 lw          20   0 8377M 5161M 4691M R 100.2  4.2  0:55.01
   2594 lw          20   0 8377M 5161M 4691M R 100.2  4.2  0:55.01
   2598 lw          20   0 8377M 5161M 4691M R 100.2  4.2  0:55.01
F1Help  F2Setup F3SearchF4FilterF5Tree  F6SortByF7Nice -F8Nice +F9Kill  F10Quit
[0] 0:llama.cpp/build/bin/llama-cli*                    "OpenCloudOS-riscv64" 13:14 14-Feb-25
```