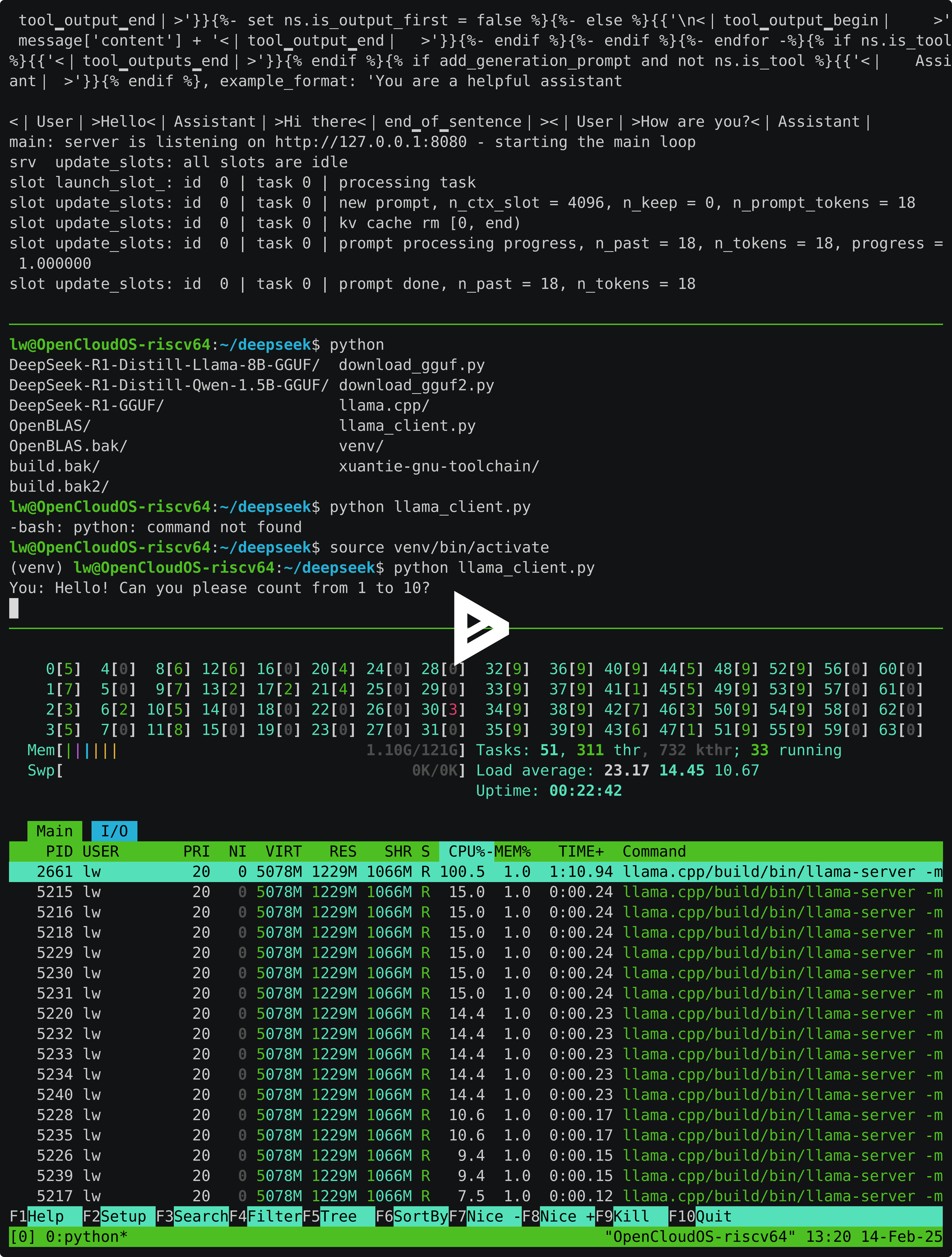

```
 tool_output_end｜>'}}{%- set ns.is_output_first = false %}{%- else %}{{'\n<｜tool_output_begin｜    >'
 message['content'] + '<｜tool_output_end｜  >'}}{%- endif %}{%- endif %}{%- endfor -%}{% if ns.is_tool
%}{{'<｜tool_outputs_end｜>'}}{% endif %}{% if add_generation_prompt and not ns.is_tool %}{{'<｜   Assi
ant｜ >'}}{% endif %}, example_format: 'You are a helpful assistant

<｜User｜>Hello<｜Assistant｜>Hi there<｜end_of_sentence｜><｜User｜>How are you?<｜Assistant｜
main: server is listening on http://127.0.0.1:8080 - starting the main loop
srv  update_slots: all slots are idle
slot launch_slot_: id  0 | task 0 | processing task
slot update_slots: id  0 | task 0 | new prompt, n_ctx_slot = 4096, n_keep = 0, n_prompt_tokens = 18
slot update_slots: id  0 | task 0 | kv cache rm [0, end)
slot update_slots: id  0 | task 0 | prompt processing progress, n_past = 18, n_tokens = 18, progress =
 1.000000
slot update_slots: id  0 | task 0 | prompt done, n_past = 18, n_tokens = 18
```

```
lw@OpenCloudOS-riscv64:~/deepseek$ python
DeepSeek-R1-Distill-Llama-8B-GGUF/  download_gguf.py
DeepSeek-R1-Distill-Qwen-1.5B-GGUF/ download_gguf2.py
DeepSeek-R1-GGUF/                   llama.cpp/
OpenBLAS/                           llama_client.py
OpenBLAS.bak/                       venv/
build.bak/                          xuantie-gnu-toolchain/
build.bak2/
lw@OpenCloudOS-riscv64:~/deepseek$ python llama_client.py
-bash: python: command not found
lw@OpenCloudOS-riscv64:~/deepseek$ source venv/bin/activate
(venv) lw@OpenCloudOS-riscv64:~/deepseek$ python llama_client.py
You: Hello! Can you please count from 1 to 10?
█
```

```
    0[5]  4[0]  8[6] 12[6] 16[0] 20[4] 24[0] 28[0]  32[9]  36[9] 40[9] 44[5] 48[9] 52[9] 56[0] 60[0]
    1[7]  5[0]  9[7] 13[2] 17[2] 21[4] 25[0] 29[0]  33[9]  37[9] 41[1] 45[5] 49[9] 53[9] 57[0] 61[0]
    2[3]  6[2] 10[5] 14[0] 18[0] 22[0] 26[0] 30[3]  34[9]  38[9] 42[7] 46[3] 50[9] 54[9] 58[0] 62[0]
    3[5]  7[0] 11[8] 15[0] 19[0] 23[0] 27[0] 31[0]  35[9]  39[9] 43[6] 47[1] 51[9] 55[9] 59[0] 63[0]
  Mem[||||||                            1.10G/121G] Tasks: 51, 311 thr, 732 kthr; 33 running
  Swp[                                       0K/0K] Load average: 23.17 14.45 10.67
                                                    Uptime: 00:22:42

  Main  I/O
    PID USER       PRI  NI  VIRT   RES   SHR S  CPU%-MEM%   TIME+  Command
   2661 lw          20   0 5078M 1229M 1066M R 100.5  1.0  1:10.94 llama.cpp/build/bin/llama-server -m
   5215 lw          20   0 5078M 1229M 1066M R  15.0  1.0  0:00.24 llama.cpp/build/bin/llama-server -m
   5216 lw          20   0 5078M 1229M 1066M R  15.0  1.0  0:00.24 llama.cpp/build/bin/llama-server -m
   5218 lw          20   0 5078M 1229M 1066M R  15.0  1.0  0:00.24 llama.cpp/build/bin/llama-server -m
   5229 lw          20   0 5078M 1229M 1066M R  15.0  1.0  0:00.24 llama.cpp/build/bin/llama-server -m
   5230 lw          20   0 5078M 1229M 1066M R  15.0  1.0  0:00.24 llama.cpp/build/bin/llama-server -m
   5231 lw          20   0 5078M 1229M 1066M R  15.0  1.0  0:00.24 llama.cpp/build/bin/llama-server -m
   5220 lw          20   0 5078M 1229M 1066M R  14.4  1.0  0:00.23 llama.cpp/build/bin/llama-server -m
   5232 lw          20   0 5078M 1229M 1066M R  14.4  1.0  0:00.23 llama.cpp/build/bin/llama-server -m
   5233 lw          20   0 5078M 1229M 1066M R  14.4  1.0  0:00.23 llama.cpp/build/bin/llama-server -m
   5234 lw          20   0 5078M 1229M 1066M R  14.4  1.0  0:00.23 llama.cpp/build/bin/llama-server -m
   5240 lw          20   0 5078M 1229M 1066M R  14.4  1.0  0:00.23 llama.cpp/build/bin/llama-server -m
   5228 lw          20   0 5078M 1229M 1066M R  10.6  1.0  0:00.17 llama.cpp/build/bin/llama-server -m
   5235 lw          20   0 5078M 1229M 1066M R  10.6  1.0  0:00.17 llama.cpp/build/bin/llama-server -m
   5226 lw          20   0 5078M 1229M 1066M R   9.4  1.0  0:00.15 llama.cpp/build/bin/llama-server -m
   5239 lw          20   0 5078M 1229M 1066M R   9.4  1.0  0:00.15 llama.cpp/build/bin/llama-server -m
   5217 lw          20   0 5078M 1229M 1066M R   7.5  1.0  0:00.12 llama.cpp/build/bin/llama-server -m
F1Help  F2Setup  F3SearchF4FilterF5Tree  F6SortByF7Nice -F8Nice +F9Kill  F10Quit
[0] 0:python*                                          "OpenCloudOS-riscv64" 13:20 14-Feb-25
```