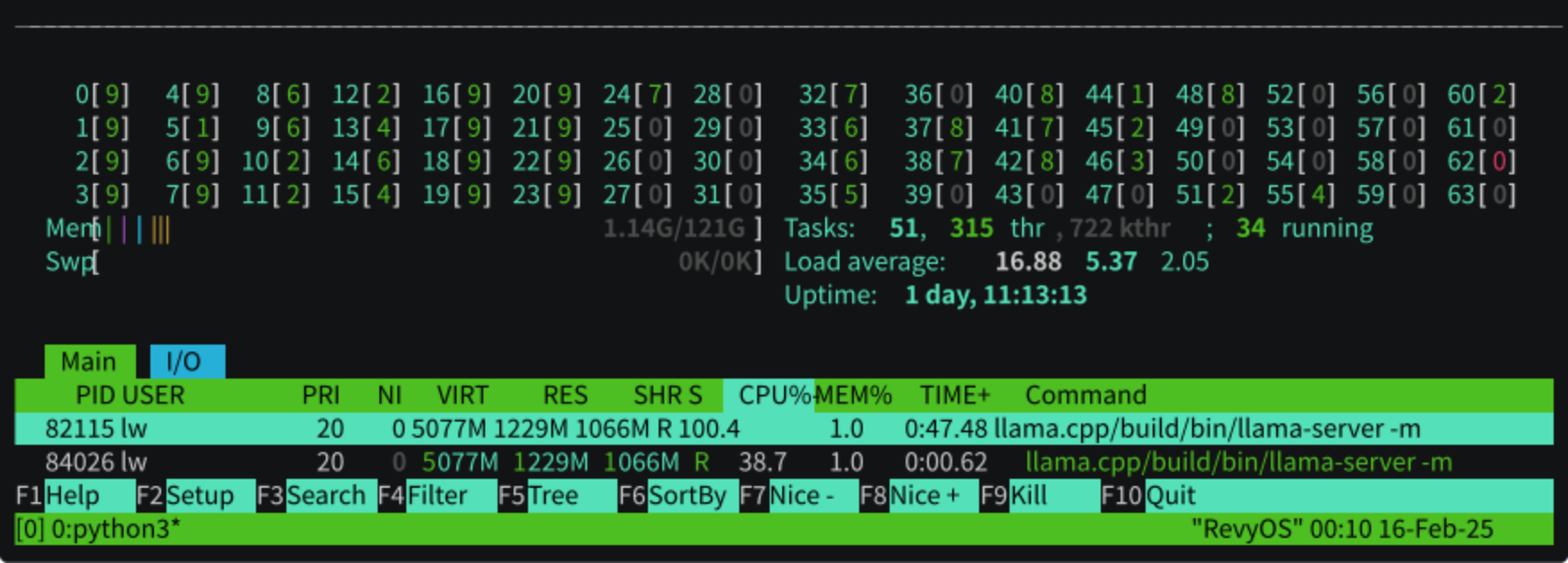
```
  File "/home/lw/deepseek/venv/lib/python3.11/site-packages/requests/adapters.py", line 667, in send
    resp = conn.urlopen(
           ^^^^^^^^^^^^^
  File "/home/lw/deepseek/venv/lib/python3.11/site-packages/urllib3/connectionpool.py", line 787, in u
rlopen
    response = self._make_request(
               ^^^^^^^^^^^^^^^^^^^
  File "/home/lw/deepseek/venv/lib/python3.11/site-packages/urllib3/connectionpool.py", line 534, in _
make_request
    response = conn.getresponse()
               ^^^^^^^^^^^^^^^^^^
  File "/home/lw/deepseek/venv/lib/python3.11/site-packages/urllib3/connection.py", line 516, in getre
sponse
    httplib_response = super().getresponse()
                       ^^^^^^^^^^^^^^^^^^^^^
  File "/usr/lib/python3.11/http/client.py", line 1378, in getresponse
    response.begin()
  File "/usr/lib/python3.11/http/client.py", line 318, in begin
    version, status, reason = self._read_status()
                              ^^^^^^^^^^^^^^^^^^^
  File "/usr/lib/python3.11/http/client.py", line 279, in _read_status
    line = str(self.fp.readline(_MAXLINE + 1), "iso-8859-1")
               ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
  File "/usr/lib/python3.11/socket.py", line 706, in readinto
    return self._sock.recv_into(b)
           ^^^^^^^^^^^^^^^^^^^^^^^
KeyboardInterrupt

(venv)  lw@RevyOS~/deepseek $ python3 ./llama_client.py
You: Hello! Who are you?
```

```
 | tool _output _end | >'}}{%- set ns.is_output_first = false %}{%- else %}{{'\n<                    | tool _output _begin | >' + me
ssage['content'] + '<          | tool _output _end | >'}}{%- endif %}{%- endif %}{%- endfor -%}{% if ns.is_tool %}{{'<
 | tool _outputs _end | >'}}{% endif %}{% if add_generation_prompt and not ns.is_tool %}{{'<                    | Assistant | >'}
% 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 = 10
slot update_slots: id  0 | task 0 | kv cache rm [0, end)
slot update_slots: id  0 | task 0 | prompt processing progress, n_past = 10, n_tokens = 10, progress =
 1.000000
slot update_slots: id  0 | task 0 | prompt done, n_past = 10, n_tokens = 10
```


0[9] 4[9] 8[6] 12[2] 16[9] 20[9] 24[7] 28[0] 32[7] 36[0] 40[8] 44[1] 48[8] 52[0] 56[0] 60[2]
1[9] 5[1] 9[6] 13[4] 17[9] 21[9] 25[0] 29[0] 33[6] 37[8] 41[7] 45[2] 49[0] 53[0] 57[0] 61[0]
2[9] 6[9] 10[2] 14[6] 18[9] 22[9] 26[0] 30[0] 34[6] 38[7] 42[8] 46[3] 50[0] 54[0] 58[0] 62[0]
3[9] 7[9] 11[2] 15[4] 19[9] 23[9] 27[0] 31[0] 35[5] 39[0] 43[0] 47[0] 51[2] 55[4] 59[0] 63[0]
Mem[1.14G/121G] Tasks: 51, 315 thr , 722 kthr ; 34 running
Swp[0K/0K] Load average: 16.88 5.37 2.05
Uptime: 1 day, 11:13:13
Main I/O
PID USER PRI NI VIRT RES SHR S CPU% MEM% TIME+ Command
82115 lw 20 0 5077M 1229M 1066M R 100.4 1.0 0:47.48 llama.cpp/build/bin/llama-server -m
84026 lw 20 0 5077M 1229M 1066M R 38.7 1.0 0:00.62 llama.cpp/build/bin/llama-server -m
F1Help F2Setup F3Search F4Filter F5Tree F6SortBy F7Nice - F8Nice + F9Kill F10Quit
[0] 0:python3*
"RevyOS" 00:10 16-Feb-25
```
```