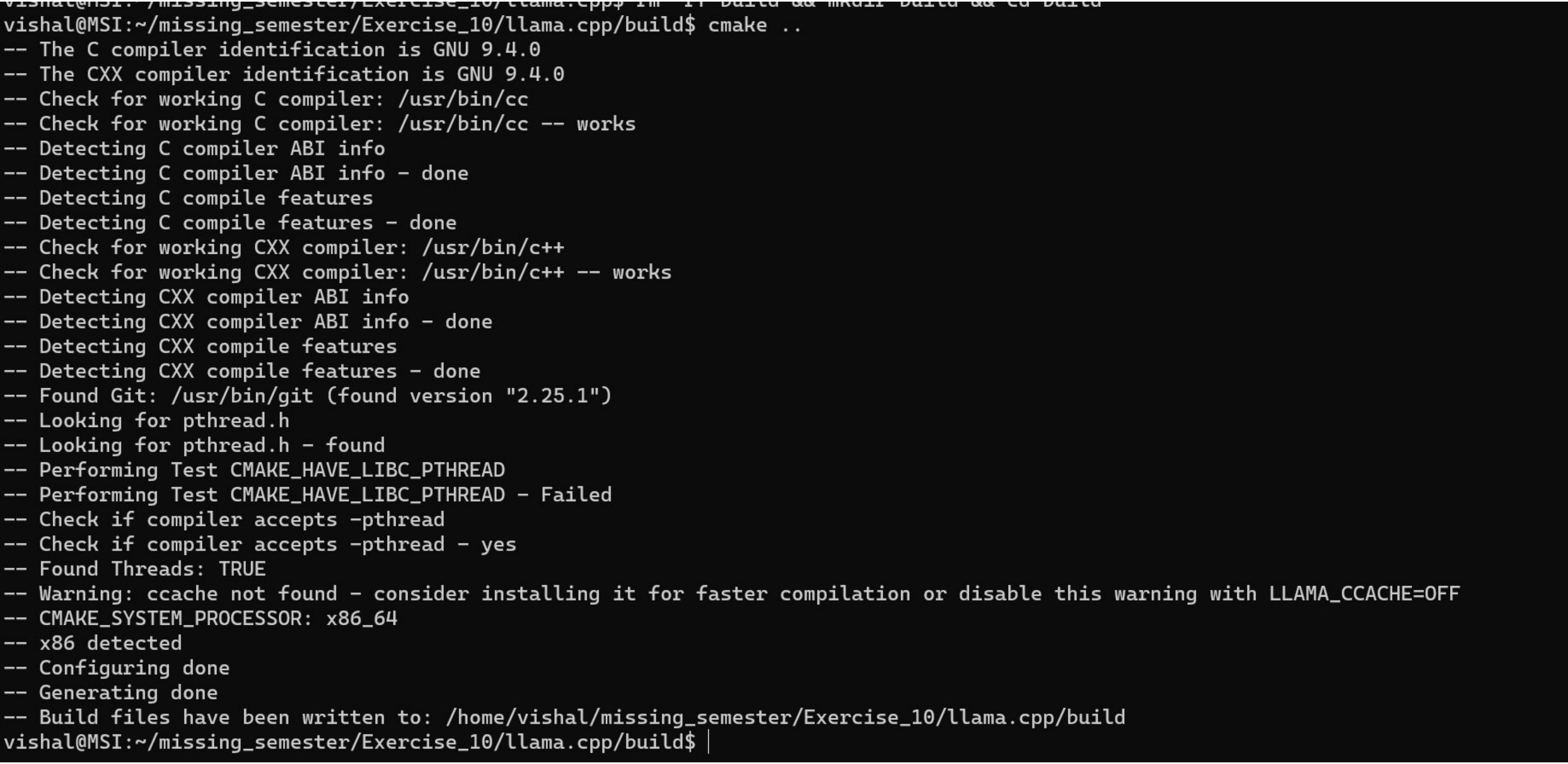
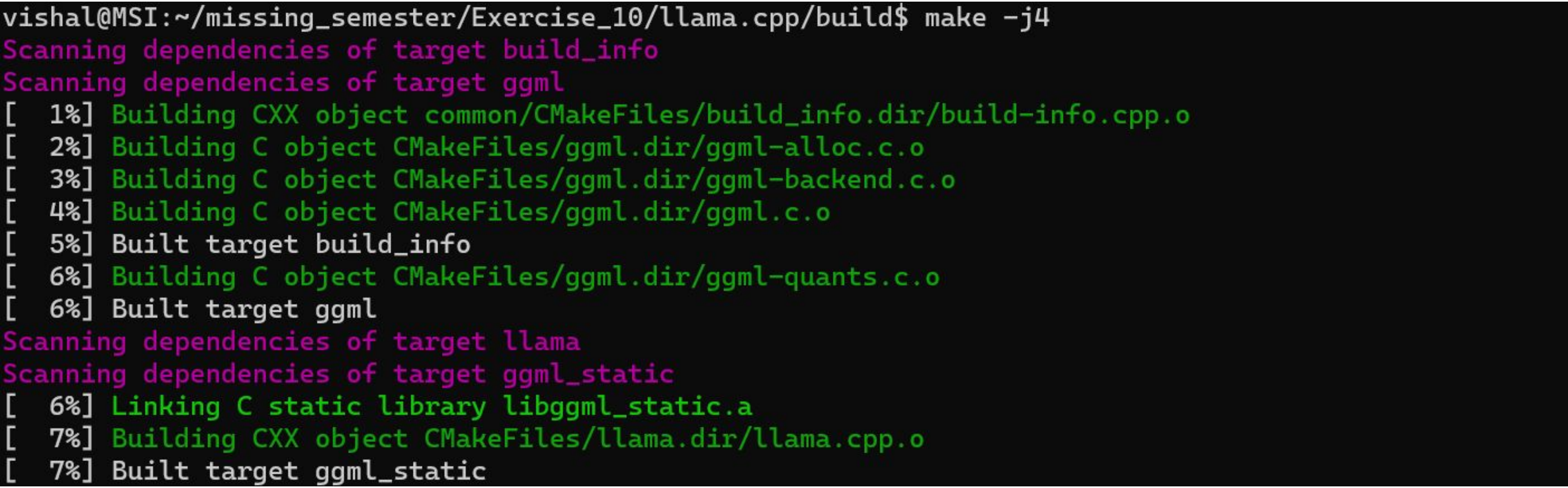
Server Setup

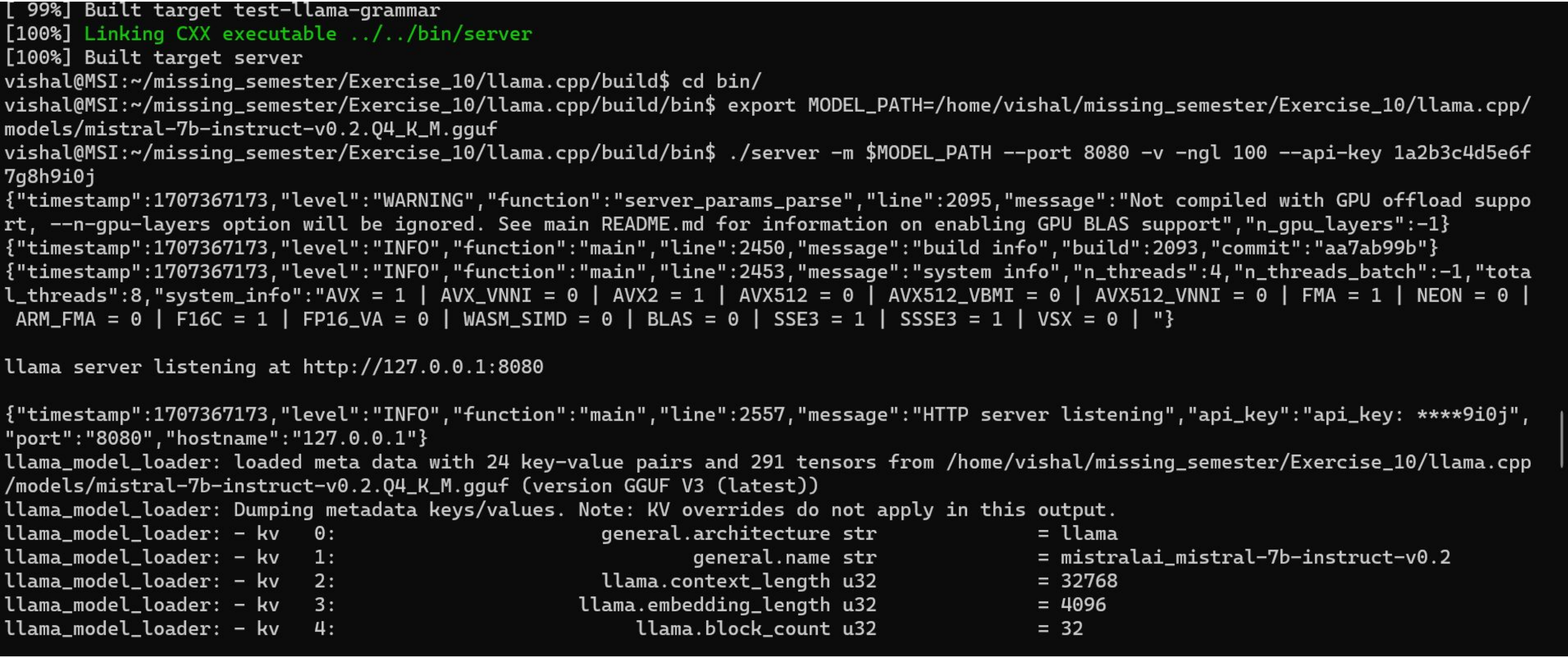
MODEL NAME - mistral-7b-instruct-v0.2.Q4_K_M.gguf



```
vishal@MSI:~/missing_semester/Exercise_10/llama.cpp$ rm -rf build && mkdir build && cd build
vishal@MSI:~/missing_semester/Exercise_10/llama.cpp/build$ cmake ..
-- The C compiler identification is GNU 9.4.0
-- The CXX compiler identification is GNU 9.4.0
-- Check for working C compiler: /usr/bin/cc
-- Check for working C compiler: /usr/bin/cc -- works
-- Detecting C compiler ABI info
-- Detecting C compiler ABI info - done
-- Detecting C compile features
-- Detecting C compile features - done
-- Check for working CXX compiler: /usr/bin/c++
-- Check for working CXX compiler: /usr/bin/c++ -- works
-- Detecting CXX compiler ABI info
-- Detecting CXX compiler ABI info - done
-- Detecting CXX compile features
-- Detecting CXX compile features - done
-- Found Git: /usr/bin/git (found version "2.25.1")
-- Looking for pthread.h
-- Looking for pthread.h - found
-- Performing Test CMAKE_HAVE_LIBC_PTHREAD
-- Performing Test CMAKE_HAVE_LIBC_PTHREAD - Failed
-- Check if compiler accepts -pthread
-- Check if compiler accepts -pthread - yes
-- Found Threads: TRUE
-- Warning: ccache not found - consider installing it for faster compilation or disable this warning with LLAMA_CCACHE=OFF
-- CMAKE_SYSTEM_PROCESSOR: x86_64
-- x86 detected
-- Configuring done
-- Generating done
-- Build files have been written to: /home/vishal/missing_semester/Exercise_10/llama.cpp/build
vishal@MSI:~/missing_semester/Exercise_10/llama.cpp/build$
```



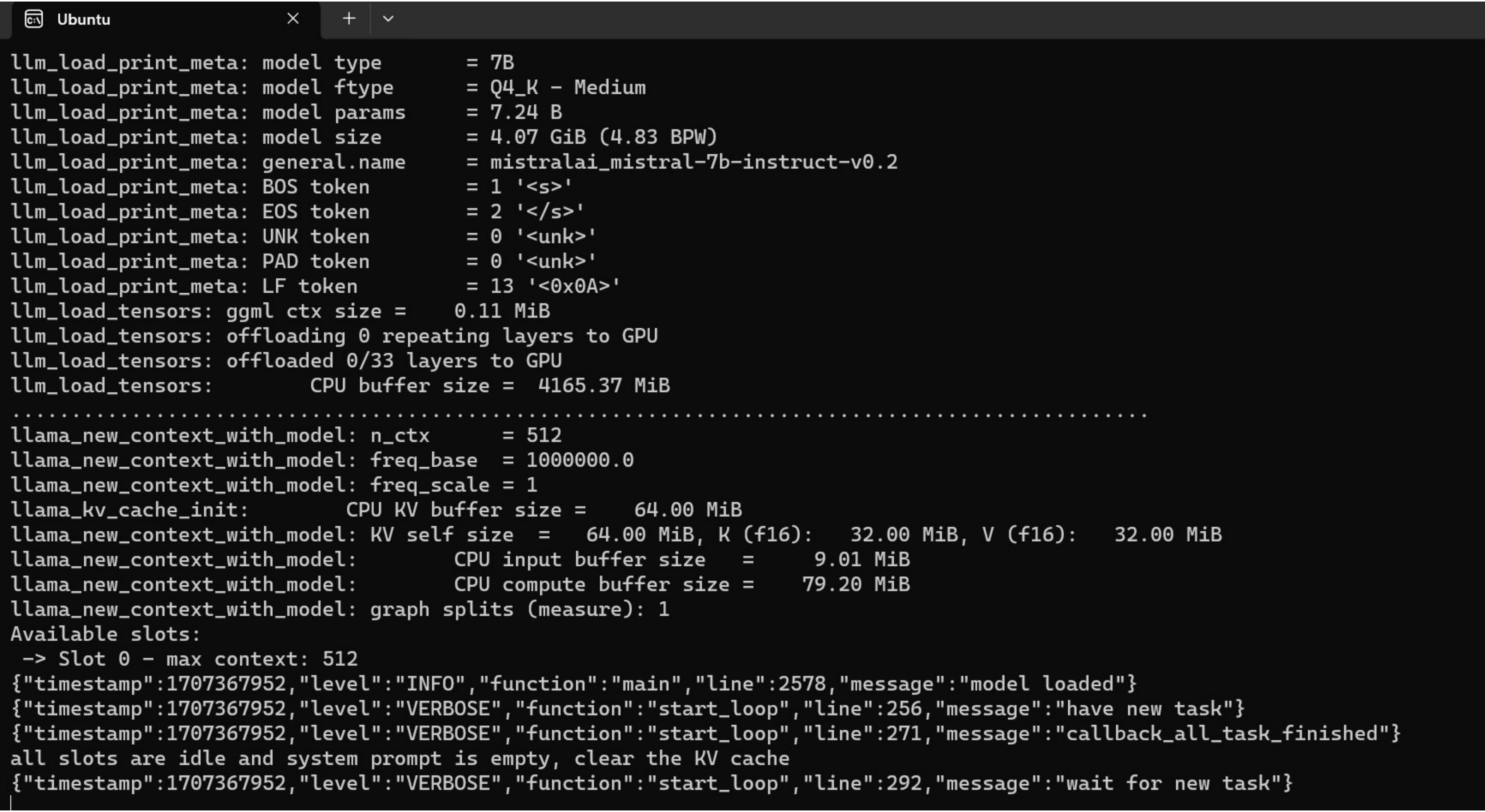
```
vishal@MSI:~/missing_semester/Exercise_10/llama.cpp/build$ make -j4
Scanning dependencies of target build_info
Scanning dependencies of target ggml
[  1%] Building CXX object common/CMakeFiles/build_info.dir/build-info.cpp.o
[  2%] Building C object CMakeFiles/ggml.dir/ggml-alloc.c.o
[  3%] Building C object CMakeFiles/ggml.dir/ggml-backend.c.o
[  4%] Building C object CMakeFiles/ggml.dir/ggml.c.o
[  5%] Built target build_info
[  6%] Building C object CMakeFiles/ggml.dir/ggml-quants.c.o
[  6%] Built target ggml
Scanning dependencies of target llama
Scanning dependencies of target ggml_static
[  6%] Linking C static library libggml_static.a
[  7%] Building CXX object CMakeFiles/llama.dir/llama.cpp.o
[  7%] Built target ggml_static
```



```
[ 99%] Built target test-llama-grammar
[100%] Linking CXX executable ../../bin/server
[100%] Built target server
vishal@MSI:~/missing_semester/Exercise_10/llama.cpp/build$ cd bin/
vishal@MSI:~/missing_semester/Exercise_10/llama.cpp/build/bin$ export MODEL_PATH=/home/vishal/missing_semester/Exercise_10/llama.cpp/
models/mistral-7b-instruct-v0.2.Q4_K_M.gguf
vishal@MSI:~/missing_semester/Exercise_10/llama.cpp/build/bin$ ./server -m $MODEL_PATH --port 8080 -v -ngl 100 --api-key 1a2b3c4d5e6f
7g8h9i0j
{"timestamp":1707367173,"level":"WARNING","function":"server_params_parse","line":2095,"message":"Not compiled with GPU offload suppo
rt, --n-gpu-layers option will be ignored. See main README.md for information on enabling GPU BLAS support","n_gpu_layers":-1}
{"timestamp":1707367173,"level":"INFO","function":"main","line":2450,"message":"build info","build":2093,"commit":"aa7ab99b"}
{"timestamp":1707367173,"level":"INFO","function":"main","line":2453,"message":"system info","n_threads":4,"n_threads_batch":-1,"tota
l_threads":8,"system_info":"AVX = 1 | AVX_VNNI = 0 | AVX2 = 1 | AVX512 = 0 | AVX512_VBMI = 0 | AVX512_VNNI = 0 | FMA = 1 | NEON = 0 |
 ARM_FMA = 0 | F16C = 1 | FP16_VA = 0 | WASM_SIMD = 0 | BLAS = 0 | SSE3 = 1 | SSSE3 = 1 | VSX = 0 | "}

llama server listening at http://127.0.0.1:8080

{"timestamp":1707367173,"level":"INFO","function":"main","line":2557,"message":"HTTP server listening","api_key":"api_key: ****9i0j",
"port":"8080","hostname":"127.0.0.1"}
llama_model_loader: loaded meta data with 24 key-value pairs and 291 tensors from /home/vishal/missing_semester/Exercise_10/llama.cpp
/models/mistral-7b-instruct-v0.2.Q4_K_M.gguf (version GGUF V3 (latest))
llama_model_loader: Dumping metadata keys/values. Note: KV overrides do not apply in this output.
llama_model_loader: - kv   0:                       general.architecture str              = llama
llama_model_loader: - kv   1:                               general.name str              = mistralai_mistral-7b-instruct-v0.2
llama_model_loader: - kv   2:                       llama.context_length u32              = 32768
llama_model_loader: - kv   3:                     llama.embedding_length u32              = 4096
llama_model_loader: - kv   4:                          llama.block_count u32              = 32
```



```
llm_load_print_meta: model type       = 7B
llm_load_print_meta: model ftype      = Q4_K - Medium
llm_load_print_meta: model params     = 7.24 B
llm_load_print_meta: model size       = 4.07 GiB (4.83 BPW)
llm_load_print_meta: general.name     = mistralai_mistral-7b-instruct-v0.2
llm_load_print_meta: BOS token        = 1 '<s>'
llm_load_print_meta: EOS token        = 2 '</s>'
llm_load_print_meta: UNK token        = 0 '<unk>'
llm_load_print_meta: PAD token        = 0 '<unk>'
llm_load_print_meta: LF token         = 13 '<0x0A>'
llm_load_tensors: ggml ctx size =    0.11 MiB
llm_load_tensors: offloading 0 repeating layers to GPU
llm_load_tensors: offloaded 0/33 layers to GPU
llm_load_tensors:        CPU buffer size =  4165.37 MiB
..................................................................................................
llama_new_context_with_model: n_ctx      = 512
llama_new_context_with_model: freq_base  = 1000000.0
llama_new_context_with_model: freq_scale = 1
llama_kv_cache_init:        CPU KV buffer size =    64.00 MiB
llama_new_context_with_model: KV self size  =   64.00 MiB, K (f16):   32.00 MiB, V (f16):   32.00 MiB
llama_new_context_with_model:        CPU input buffer size   =     9.01 MiB
llama_new_context_with_model:        CPU compute buffer size =    79.20 MiB
llama_new_context_with_model: graph splits (measure): 1
Available slots:
 -> Slot 0 - max context: 512
{"timestamp":1707367952,"level":"INFO","function":"main","line":2578,"message":"model loaded"}
{"timestamp":1707367952,"level":"VERBOSE","function":"start_loop","line":256,"message":"have new task"}
{"timestamp":1707367952,"level":"VERBOSE","function":"start_loop","line":271,"message":"callback_all_task_finished"}
all slots are idle and system prompt is empty, clear the KV cache
{"timestamp":1707367952,"level":"VERBOSE","function":"start_loop","line":292,"message":"wait for new task"}
```