CS6886W – System Engineering for Deep Learning
Assignment 2: Performance Analysis of GPT-2

Submitted by: Adamaya Sharma

Course: CS6886W – System Engineering for DL

Roll No: CS24M501

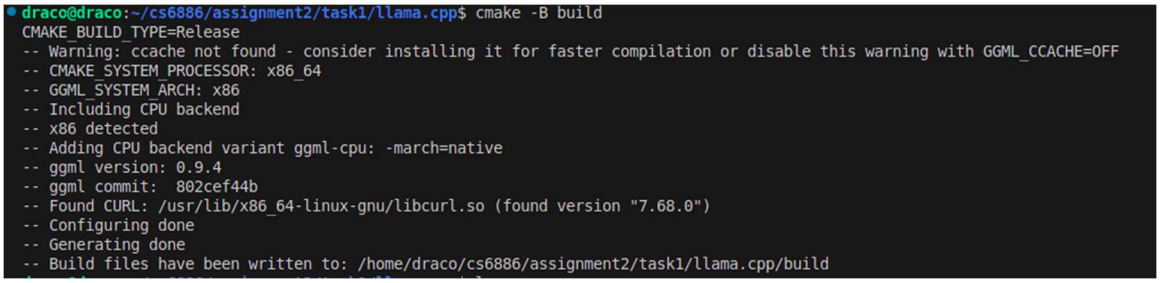
Github Link: https://github.com/Adamaya/cs6886w-assignment2

Task 1 – Installation of llama.cpp

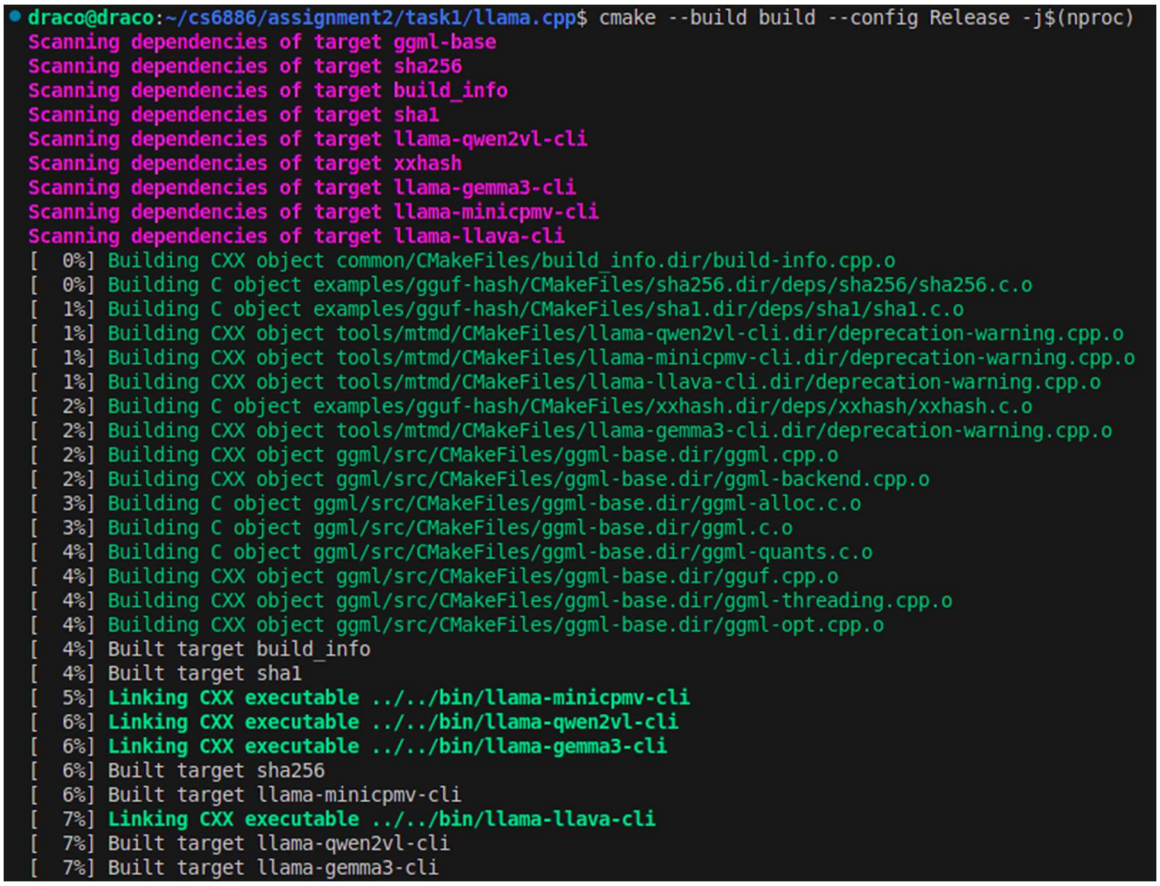
Objective: Build and install llama.cpp from source.

Steps Followed:

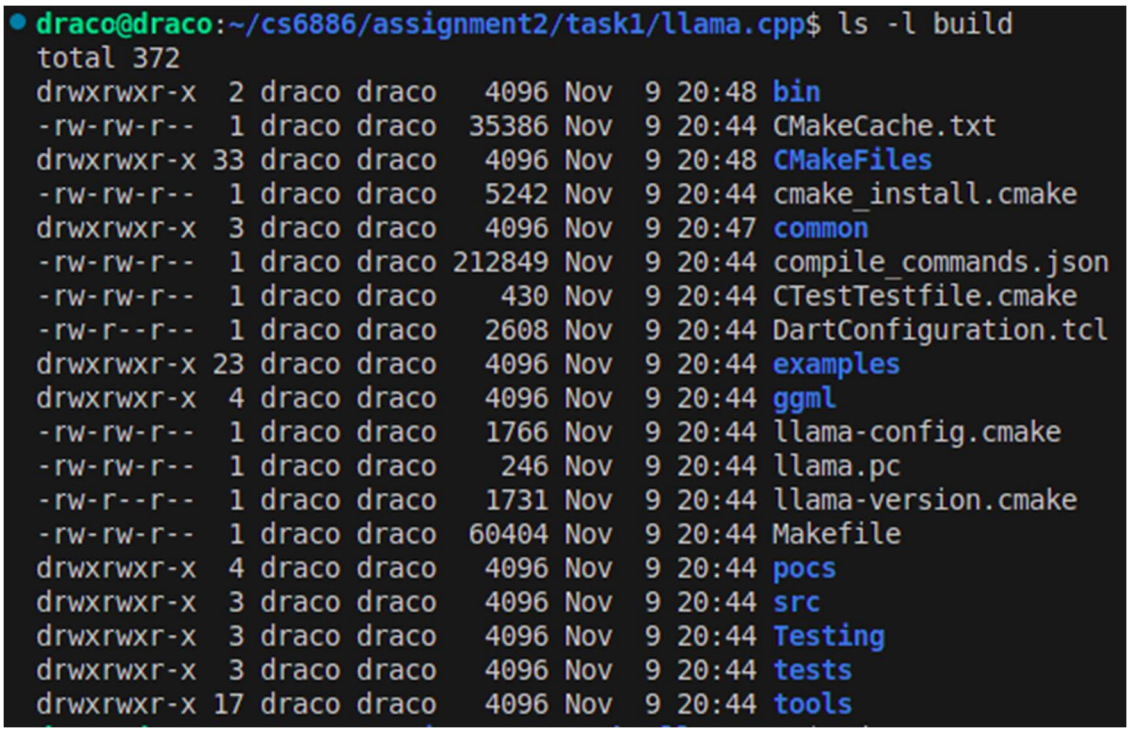
- Ran cmake to configure CPU backend
  - CMake setup for CPU backend on x86-64 architecture.



- Compiled executables using parallel build.
  - Parallel build completed with multiple CLI executables



- Verified generated build directory and executables.
  - Listing of generated files confirming successful installation



System environment:

- **CPU** : 12th Gen Intel Core i7-1270P (16 threads, 12 cores)
- **Architecture** : x86-64 (Little Endian)
- **Operating System** : Ubuntu 20.04.1 Linux Kernel version 5.15.0-149-generic (SMP, April 16 2025 build)

Results: Successful build confirmation with GGML v0.9.4.

Task 2 – Setting Up GPT-2 Medium Model

Objective: Download GPT-2 Medium and convert it to GGUF format.

Steps Followed:

- Cloned GPT-2 Medium from Hugging Face.
    - Successful clone of GPT-2 Medium from Hugging Face

```
draco@draco:~/cs6886/assignment2$ git clone https://huggingface.co/openai-community/gpt2-medium
Cloning into 'gpt2-medium'...
remote: Enumerating objects: 76, done.
remote: Total 76 (delta 0), reused 0 (delta 0), pack-reused 76 (from 1)
Unpacking objects: 100% (76/76), 1.65 MiB | 1.95 MiB/s, done.
```

    - Clean directory organization under task2/

```
draco@draco:~/cs6886/assignment2$ mkdir task2
draco@draco:~/cs6886/assignment2$ git clone https://huggingface.co/openai-community/gpt2-medium
Cloning into 'gpt2-medium'...
remote: Enumerating objects: 76, done.
remote: Total 76 (delta 0), reused 0 (delta 0), pack-reused 76 (from 1)
Unpacking objects: 100% (76/76), 1.65 MiB | 1.95 MiB/s, done.
draco@draco:~/cs6886/assignment2$ mv gpt2-medium/ task2/
```

- Converted model using convert_hf_to_gguf.py.

```
draco@draco:~/cs6886/assignment2/task1/llama.cpp$ python3 convert_hf_to_gguf.py ../../task2/gpt2-medium --outfile gpt2-medium.gguf
INFO:hf-to-gguf:Loading model: gpt2-medium
INFO:hf-to-gguf:Model architecture: GPT2LMHeadModel
INFO:hf-to-gguf:gguf: indexing model part 'model.safetensors'
INFO:gguf.gguf_writer:gguf: This GGUF file is for Little Endian only
INFO:hf-to-gguf:Exporting model...
INFO:hf-to-gguf:blk.0.attn_qkv.bias,       torch.float32 --> F32, shape = {3072}
INFO:hf-to-gguf:blk.0.attn_qkv.weight,     torch.float32 --> F16, shape = {1024, 3072}
INFO:gguf.gguf_writer:Writing the following files:
INFO:gguf.gguf_writer:gpt2-medium.gguf: n_tensors = 292, total_size = 712.4M
Writing:   0%|                                | 0.00/712M [00:00<?, ?byte/s]/home/draco/cs6886/assignment2/task1/llama.cpp/gguf-py/gguf/lazy.py:222: RuntimeWarning: overflow encountered in cast
  return type(self)(meta=meta, args=full_args, kwargs=kwargs, func=(lambda a, *args, **kwargs: a.astype(*args, **kwargs)))
Writing: 100%|████████████████████████████████| 712M/712M [00:22<00:00, 31.4Mbyte/s]
INFO:hf-to-gguf:Model successfully exported to gpt2-medium.gguf
```

- Ran llama-bench sanity test.
    - Validated GGUF conversion with ≈ 49.65 t/s throughput

```
draco@draco:~/cs6886/assignment2/task1/llama.cpp$ ./build/bin/llama-bench -m gpt2-medium.gguf -p 0 -n 256
| model                          |       size |     params | backend    | threads |           test |                  t/s |
| ------------------------------ | ---------: | ---------: | ---------- | ------: | -------------: | -------------------: |
| gpt2 0.4B F16                  | 679.38 MiB |   354.82 M | CPU        |       4 |          tg256 |         49.65 ± 3.01 |

build: 802cef44b (7001)
```

Results: Model successfully benchmarked at 49.65 tokens/s.

Task 3 – Naive Execution (No Parallelism)

Objective: Run scalar build without SIMD or threading.

Steps Followed:

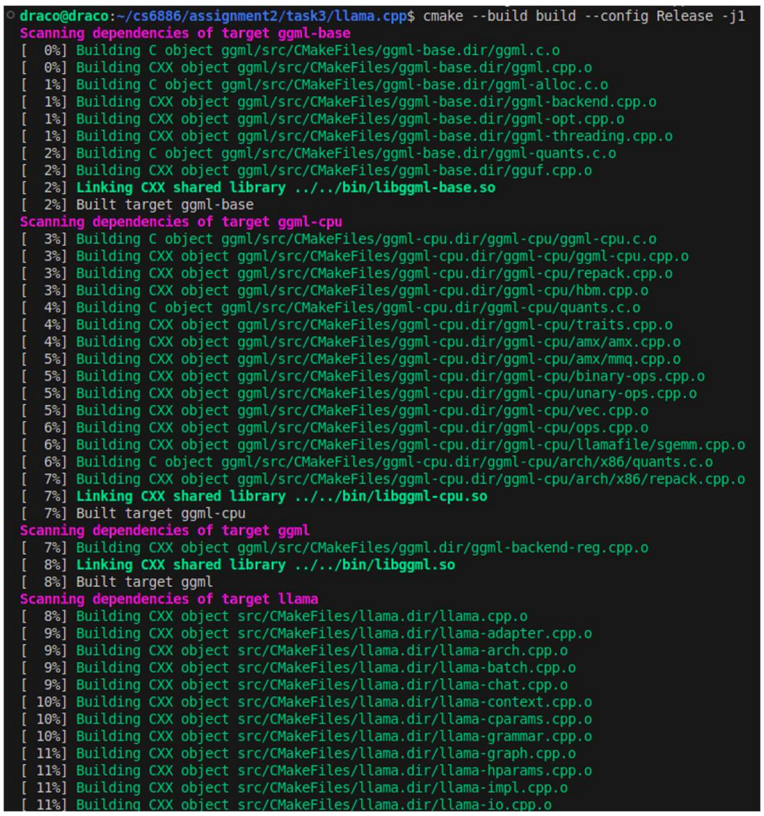
- Disabled all AVX/SSE/FMA flags.
    - Fresh clone of llama.cpp

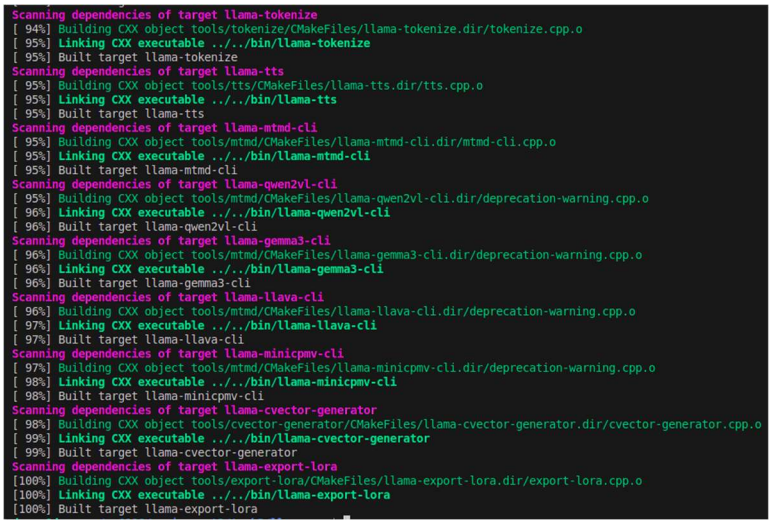
```
draco@draco:~/cs6886/assignment2/task3$ git clone https://github.com/ggml-org/llama.cpp.git
Cloning into 'llama.cpp'...
remote: Enumerating objects: 67278, done.
remote: Counting objects: 100% (28/28), done.
remote: Compressing objects: 100% (20/20), done.
remote: Total 67278 (delta 11), reused 11 (delta 8), pack-reused 67250 (from 2)
Receiving objects: 100% (67278/67278), 195.35 MiB | 633.00 KiB/s, done.
Resolving deltas: 100% (48578/48578), done.
```

    - SIMD, FMA, and AVX paths disable

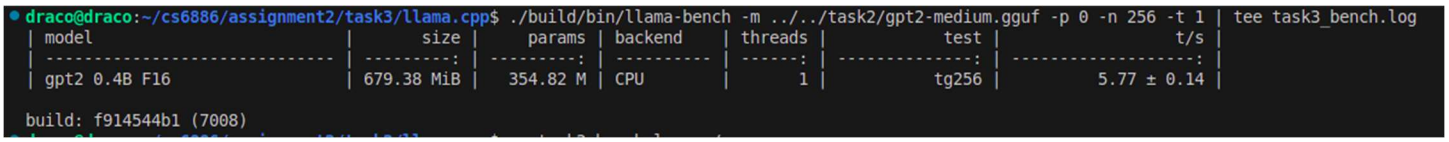
```
draco@draco:~/cs6886/assignment2/task3$ cd llama.cpp/
draco@draco:~/cs6886/assignment2/task3/llama.cpp$ cmake -B build  -DGGML_CPU_GENERIC=ON  -DGGML_NATIVE=OFF  -DGGML_AVX=OFF  -DGGML_AVX2=OFF  -DGGML_AVX512=OFF  -DGGML_SSE42=OFF  -DGGM
L_F16C=OFF  -DGGML_FMA=OFF
-- The C compiler identification 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
CMAKE_BUILD_TYPE=Release
-- Found Git: /usr/bin/git (found version "2.25.1")
-- The ASM compiler identification is GNU
-- Found assembler: /usr/bin/cc
-- 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 GGML_CCACHE=OFF
-- CMAKE_SYSTEM_PROCESSOR: x86_64
-- GGML_SYSTEM_ARCH: x86
-- Including CPU backend
-- Found OpenMP_C: -fopenmp (found version "4.5")
-- Found OpenMP_CXX: -fopenmp (found version "4.5")
-- Found OpenMP: TRUE (found version "4.5")
-- x86 detected
-- Adding CPU backend variant ggml-cpu: -mbmi2 GGML_BMI2
-- ggml version: 0.9.4
-- ggml commit:  f914544b1
-- Found CURL: /usr/lib/x86_64-linux-gnu/libcurl.so (found version "7.68.0")
-- Configuring done
-- Generating done
-- Build files have been written to: /home/draco/cs6886/assignment2/task3/llama.cpp/build
```

    - Successful scalar build of CLI binarie





- Ran single-thread benchmark.



Results: 5.77 tokens/s throughput.

Model	Size	Params	Backend	Threads	Test	Throughput (t/s)
GPT-2 0.4B F16	679.38 MiB	354.82 M	CPU	1	tg256	**5.77 ± 0.14**

Task 4 – Default Execution (Single Thread)

Objective: Enable CPU-native vectorization.

Steps Followed:

- Default cmake build with SIMD enabled.
    - Default CMake build with full CPU optimizations

```
draco@draco:~/cs6886/assignment2/task4/llama.cpp$ cmake -B build
-- 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
CMAKE_BUILD_TYPE=Release
-- Found Git: /usr/bin/git (found version "2.25.1")
-- The ASM compiler identification is GNU
-- Found assembler: /usr/bin/cc
-- 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 GGML_CCACHE=OFF
-- CMAKE_SYSTEM_PROCESSOR: x86_64
-- GGML_SYSTEM_ARCH: x86
-- Including CPU backend
-- Found OpenMP_C: -fopenmp (found version "4.5")
-- Found OpenMP_CXX: -fopenmp (found version "4.5")
-- Found OpenMP: TRUE (found version "4.5")
-- x86 detected
-- Adding CPU backend variant ggml-cpu: -march=native
-- ggml version: 0.9.4
-- ggml commit:  df70bedda
-- Found CURL: /usr/lib/x86_64-linux-gnu/libcurl.so (found version "7.68.0")
-- Configuring done
-- Generating done
-- Build files have been written to: /home/draco/cs6886/assignment2/task4/llama.cpp/build
```

    - Successful vectorized build (SIMD enabled)

```
draco@draco:~/cs6886/assignment2/task4/llama.cpp$ cmake --build build --config Release -j$(nproc)
Scanning dependencies of target build_info
Scanning dependencies of target sha256
Scanning dependencies of target ggml-base
Scanning dependencies of target sha1
Scanning dependencies of target xxhash
Scanning dependencies of target llama-gemma3-cli
Scanning dependencies of target llama-qwen2vl-cli
Scanning dependencies of target llama-llava-cli
Scanning dependencies of target llama-minicpmv-cli
[  0%] Building C object examples/gguf-hash/CMakeFiles/sha256.dir/deps/sha256/sha256.c.o
[  0%] Building CXX object common/CMakeFiles/build_info.dir/build-info.cpp.o
[  1%] Building C object examples/gguf-hash/CMakeFiles/sha1.dir/deps/sha1/sha1.c.o
[  1%] Building CXX object tools/mtmd/CMakeFiles/llama-qwen2vl-cli.dir/deprecation-warning.cpp.o
[  1%] Building CXX object tools/mtmd/CMakeFiles/llama-minicpmv-cli.dir/deprecation-warning.cpp.o
[  1%] Building CXX object tools/mtmd/CMakeFiles/llama-llava-cli.dir/deprecation-warning.cpp.o
[  1%] Building CXX object tools/mtmd/CMakeFiles/llama-gemma3-cli.dir/deprecation-warning.cpp.o
[  1%] Building CXX object ggml/src/CMakeFiles/ggml-base.dir/ggml.cpp.o
[  2%] Building C object examples/gguf-hash/CMakeFiles/xxhash.dir/deps/xxhash/xxhash.c.o
[  1%] Building C object ggml/src/CMakeFiles/ggml-base.dir/ggml.c.o
[  3%] Building C object ggml/src/CMakeFiles/ggml-base.dir/ggml-alloc.c.o
[  3%] Building CXX object ggml/src/CMakeFiles/ggml-base.dir/ggml-opt.cpp.o
[  3%] Building CXX object ggml/src/CMakeFiles/ggml-base.dir/ggml-backend.cpp.o
[  4%] Building C object ggml/src/CMakeFiles/ggml-base.dir/ggml-quants.c.o
[  4%] Building CXX object ggml/src/CMakeFiles/ggml-base.dir/gguf.cpp.o
[  4%] Building CXX object ggml/src/CMakeFiles/ggml-base.dir/ggml-threading.cpp.o
[  4%] Built target build_info
[  4%] Built target sha1
[  5%] Linking CXX executable ../../bin/llama-qwen2vl-cli
[  5%] Built target sha256
[  6%] Linking CXX executable ../../bin/llama-minicpmv-cli
[  7%] Linking CXX executable ../../bin/llama-llava-cli
[  7%] Built target llama-qwen2vl-cli
[  7%] Linking CXX executable ../../bin/llama-gemma3-cli
[  7%] Built target llama-minicpmv-cli
[  7%] Built target llama-gemma3-cli
[  7%] Built target llama-llava-cli
[  7%] Built target xxhash
[  7%] Linking CXX shared library ../../bin/libggml-base.so
[  7%] Built target ggml-base
Scanning dependencies of target ggml-cpu
[  8%] Building C object ggml/src/CMakeFiles/ggml-cpu.dir/ggml-cpu/ggml-cpu.c.o
[  8%] Building CXX object ggml/src/CMakeFiles/ggml-cpu.dir/ggml-cpu/ggml-cpu.cpp.o
[  9%] Building C object ggml/src/CMakeFiles/ggml-cpu.dir/ggml-cpu/quants.c.o
[  9%] Building CXX object ggml/src/CMakeFiles/ggml-cpu.dir/ggml-cpu/repack.cpp.o
[  9%] Building CXX object ggml/src/CMakeFiles/ggml-cpu.dir/ggml-cpu/hbm.cpp.o
[  9%] Building CXX object ggml/src/CMakeFiles/ggml-cpu.dir/ggml-cpu/traits.cpp.o
Scanning dependencies of target llama-export-lora
[ 95%] Building CXX object tools/export-lora/CMakeFiles/llama-export-lora.dir/export-lora.cpp.o
[ 95%] Built target test-json-schema-to-grammar
[ 96%] Linking CXX executable ../../bin/llama-tokenize
[ 96%] Built target llama-tokenize
[ 96%] Linking CXX executable ../../bin/llama-quantize
Scanning dependencies of target llama-server
[ 97%] Building CXX object tools/server/CMakeFiles/llama-server.dir/server.cpp.o
[ 97%] Built target llama-quantize
[ 97%] Linking CXX executable ../../bin/llama-mtmd-cli
[ 98%] Linking CXX executable ../../bin/llama-cli
[ 98%] Linking CXX executable ../../bin/llama-export-lora
[ 98%] Built target llama-mtmd-cli
[ 98%] Built target llama-cli
[ 98%] Built target llama-export-lora
[ 99%] Linking CXX executable ../../bin/llama-cvector-generator
[ 99%] Built target llama-cvector-generator
[ 99%] Linking CXX executable ../../bin/llama-perplexity
[ 99%] Built target llama-perplexity
[ 99%] Linking CXX executable ../../bin/llama-run
[ 99%] Built target llama-run
[ 99%] Linking CXX executable ../../bin/llama-imatrix
[100%] Linking CXX executable ../bin/test-chat
[100%] Built target test-chat
[100%] Built target llama-imatrix
[100%] Linking CXX executable ../../bin/llama-bench
[100%] Built target llama-bench
[100%] Linking CXX executable ../../bin/llama-tts
[100%] Built target llama-tts
[100%] Linking CXX executable ../bin/test-backend-ops
[100%] Built target test-backend-ops
[100%] Linking CXX executable ../../bin/llama-server
[100%] Built target llama-server
```

- Single-thread benchmark.

Results: 24.35 tokens/s (~4.2× faster).

Model	Size	Params	Backend	Threads	Test	Throughput (t/s)
GPT-2 0.4B F16	679.38 MiB	354.82 M	CPU	1	tg256	**24.35 ± 0.46**

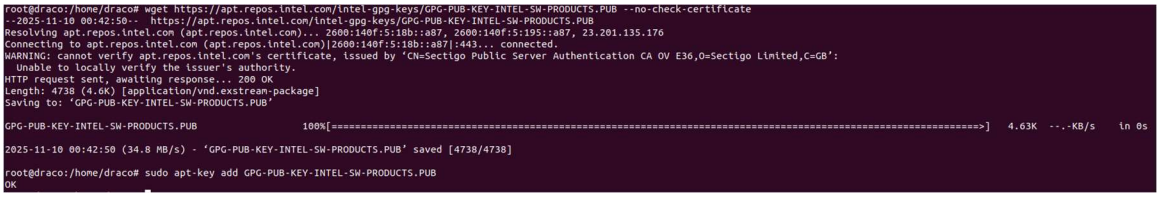
Analysis: SIMD vectorization improves arithmetic throughput.

Task 5 – Near-Optimal Execution with Intel MKL

Objective: Build llama.cpp using Intel MKL BLAS backend.

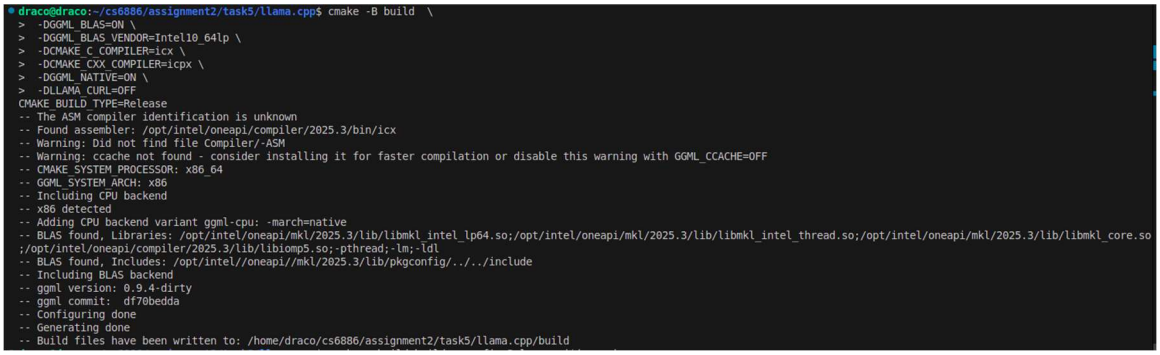
Steps Followed:

- Installed Intel oneAPI and sourced environment.
    - Verified GPG key and APT repository setup for Intel oneAPI



    - Intel MKL and BLAS detected with icx/icpx compilers

Enabled MKL and BLAS vendor flags.



Results: 30.22 tokens/s throughput, 24% faster than default.

Analysis: MKL optimizations enhance data reuse.

Configuration	SIMD/Vectorization	Library	Threads	Throughput (t/s)	Speedup vs Naive	Speedup vs Default
Task 3 – Naive	Disabled	None	1	5.77 ± 0.14	—	—
Task 4 – Default	Enabled	None	1	24.35 ± 0.46	4.2×	—
Task 5 – MKL BLAS	Enabled	Intel MKL BLAS	1	**30.22 ± 0.18**	**5.2×**	**1.24×**

Task 6 – Reporting Floating-Point Performance Counters

Objective: Collect floating-point and memory counters using perf.

Steps Followed:

- Ran perf list and filtered relevant counters.

```
draco@draco:~/cs6886/assignment2/task6$ perf list > perf_list
```

```
draco@draco:~/cs6886/assignment2/task6$ perf list | egrep -i "cpu_core/topdown|
cpu_core/cpu-cy|cpu_core/instructions|cpu_core/cache-misses|cpu_atom\cache-miss
|uncore_imc_free_running"
  cache-misses OR cpu_core/cache-misses/             [Kernel PMU event]
  cpu-cycles OR cpu_core/cpu-cycles/                 [Kernel PMU event]
  instructions OR cpu_core/instructions/             [Kernel PMU event]
  topdown-bad-spec OR cpu_core/topdown-bad-spec/     [Kernel PMU event]
  topdown-be-bound OR cpu_core/topdown-be-bound/     [Kernel PMU event]
  topdown-br-mispredict OR cpu_core/topdown-br-mispredict/ [Kernel PMU event]
  topdown-fe-bound OR cpu_core/topdown-fe-bound/     [Kernel PMU event]
  topdown-fetch-lat OR cpu_core/topdown-fetch-lat/   [Kernel PMU event]
  topdown-heavy-ops OR cpu_core/topdown-heavy-ops/   [Kernel PMU event]
  topdown-mem-bound OR cpu_core/topdown-mem-bound/   [Kernel PMU event]
  topdown-retiring OR cpu_core/topdown-retiring/     [Kernel PMU event]
  uncore_imc_free_running_0/data_read/               [Kernel PMU event]
  uncore_imc_free_running_0/data_total/              [Kernel PMU event]
  uncore_imc_free_running_0/data_write/              [Kernel PMU event]
  uncore_imc_free_running_1/data_read/               [Kernel PMU event]
  uncore_imc_free_running_1/data_total/              [Kernel PMU event]
  uncore_imc_free_running_1/data_write/              [Kernel PMU event]
```

Key Floating-Point and Memory Counters Identified

Category	Performance Counter Name	Description
Floating-Point Operations	fp_arith_inst_retired.scalar_single	Counts retired single-precision scalar floating-point operations.
	fp_arith_inst_retired.scalar_double	Retired double-precision scalar operations.
	fp_arith_inst_retired.128b_packed_single	Retired 128-bit SIMD packed single-precision operations.
	fp_arith_inst_retired.256b_packed_double	Retired 256-bit SIMD packed double-precision operations.
	fp_arith_inst_retired.vector	Total vector floating-point instructions executed.

Category	Performance Counter Name	Description
Memory and Cache Traffic	uncore_imc_free_running_0/data_read/	Total integrated memory controller (IMC) data read operations.
	uncore_imc_free_running_0/data_write/	Total IMC data writes to DRAM.
	uncore_imc_free_running_0/data_total/	Aggregated memory traffic (read + write).
	cpu_core/cache-misses	Cache misses at L1/L2/L3 levels.
	cpu_core/cache-references	Total cache references (hit + miss).
Top-Down Microarchitectural Analysis	cpu_core/topdown-retiring/	Fraction of uops that were executed successfully.
	cpu_core/topdown-heavy-ops/	Count of high-latency heavy instructions (e.g., divisions, FMAs).

Results: Identified fp_arith_inst_retired.* and uncore_imc_free_running events.
Analysis: Counters provide key input for Roofline model.

Task 7 – Performance Counters and Roofline Analysis

Objective: Analyze operation intensity and Roofline model.

Steps Followed:

• Collected perf stats for naive, default, and MKL builds.

```
draco@draco:~/cs6886/assignment2/task4/llama.cpp$ sudo perf stat   -e cpu_core/topdown-retiring/   -e cpu_core/topdown-heavy-ops/   -e cpu_core/cpu-cycles/   -e cpu_core/instructions/  -e cpu_core/cache-
misses/ -e cpu_atom/cache-misses/ -e uncore_imc_free_running_0/data_read/   -e uncore_imc_free_running_0/data_write/  ./build/bin/llama-bench  -m ../../task2/gpt2-medium/gpt2-medium.gguf -p 0 -n 256 -t 1
   2>&1 | tee perf_task4.txt
[sudo] password for draco:
| model                          |       size |     params | backend    | threads |            test |                  t/s |
| ------------------------------ | ---------: | ---------: | ---------- | ------: | --------------: | -------------------: |
| gpt2 0.4B F16                  | 679.38 MiB |   354.82 M | CPU        |       1 |           tg256 |         30.42 ± 0.67 |

build: df70bedda (7009)

 Performance counter stats for './build/bin/llama-bench -m ../../task2/gpt2-medium/gpt2-medium.gguf -p 0 -n 256 -t 1':

   <not supported>      cpu_core/topdown-retiring/
   <not supported>      cpu_core/topdown-heavy-ops/
 2,00,15,81,49,701      cpu_core/cpu-cycles/                                  (99.97%)
 2,76,85,05,26,605      cpu_core/instructions/                                (99.97%)
   13,53,98,22,983      cpu_core/cache-misses/                                (99.97%)
      20,01,63,548      cpu_atom/cache-misses/                                (0.03%)
   <not supported> MiB  uncore_imc_free_running_0/data_read/
   <not supported> MiB  uncore_imc_free_running_0/data_write/

      42.318826317 seconds time elapsed

      42.234846000 seconds user
       0.035995000 seconds sys
```

```
draco@draco:~/cs6886/assignment2/task3/llama.cpp$ sudo perf stat   -e cpu_core/topdown-retiring/   -e cpu_core/topdown-heavy-ops/   -e cpu_core/cpu-cycles/   -e cpu_core/instructions/  -e cpu_core/cache-
misses/ -e cpu_atom/cache-misses/ -e uncore_imc_free_running_0/data_read/   -e uncore_imc_free_running_0/data_write/  ./build/bin/llama-bench  -m ../../task2/gpt2-medium/gpt2-medium.gguf -p 0 -n 256 -t 1
   2>&1 | tee perf_alderlake.txt
| model                          |       size |     params | backend    | threads |            test |                  t/s |
| ------------------------------ | ---------: | ---------: | ---------- | ------: | --------------: | -------------------: |
| gpt2 0.4B F16                  | 679.38 MiB |   354.82 M | CPU        |       1 |           tg256 |          5.71 ± 0.02 |

build: f914544b1 (7008)

 Performance counter stats for './build/bin/llama-bench -m ../../task2/gpt2-medium/gpt2-medium.gguf -p 0 -n 256 -t 1':

   <not supported>      cpu_core/topdown-retiring/
   <not supported>      cpu_core/topdown-heavy-ops/
 9,83,67,91,27,745      cpu_core/cpu-cycles/                                  (99.97%)
47,29,60,03,64,128      cpu_core/instructions/                                (99.97%)
    7,71,77,39,678      cpu_core/cache-misses/                                (99.97%)
    8,09,08,32,255      cpu_atom/cache-misses/                                (0.03%)
   <not supported> MiB  uncore_imc_free_running_0/data_read/
   <not supported> MiB  uncore_imc_free_running_0/data_write/

     224.343339584 seconds time elapsed

     224.255040000 seconds user
       0.043997000 seconds sys
```

```
draco@draco:~/cs6886/assignment2/task5/llama.cpp$ source /opt/intel/oneapi/setvars.sh

:: initializing oneAPI environment ...
   bash: BASH_VERSION = 5.0.17(1)-release
   args: Using "$@" for setvars.sh arguments:
:: advisor -- latest
:: ccl -- latest
:: compiler -- latest
:: dal -- latest
:: debugger -- latest
:: dev-utilities -- latest
:: dnnl -- latest
:: dpcpp-ct -- latest
:: dpl -- latest
:: ipp -- latest
:: ippcp -- latest
:: mkl -- latest
:: mpi -- latest
:: tbb -- latest
:: umf -- latest
:: vtune -- latest
:: oneAPI environment initialized ::

draco@draco:~/cs6886/assignment2/task5/llama.cpp$ sudo env LD_LIBRARY_PATH="$LD_LIBRARY_PATH" perf stat   -e cpu_core/topdown-retiring/   -e cpu_core/topdown-heavy-ops/   -e cpu_core/cpu-cycles/   -e cpu
_core/instructions/  -e cpu_core/cache-misses/ -e cpu_atom/cache-misses/ -e uncore_imc_free_running_0/data_read/   -e uncore_imc_free_running_0/data_write/  ./build/bin/llama-bench  -m ../../task2/gpt2-m
edium/gpt2-medium.gguf -p 0 -n 256 -t 1   2>&1 | tee perf_alderlake.txt
^Cdraco@draco:~/cs6886/assignment2/task5/llama.cpp$ sudo env LD_LIBRARY_PATH="$LD_LIBRARY_PATH" perf stat   -e cpu_core/topdown-retiring/   -e cpu_core/topdown-heavy-ops/   -e cpu_core/cpu-cycles/   -e c
pu_core/instructions/  -e cpu_core/cache-misses/ -e cpu_atom/cache-misses/ -e uncore_imc_free_running_0/data_read/   -e uncore_imc_free_running_0/data_write/  ./build/bin/llama-bench  -m ../../task2/gpt2
-medium/gpt2-medium.gguf -p 0 -n 256 -t 1   2>&1 | tee perf_task5.txt
| model                          |       size |     params | backend    | threads |            test |                  t/s |
| ------------------------------ | ---------: | ---------: | ---------- | ------: | --------------: | -------------------: |
| gpt2 0.4B F16                  | 679.38 MiB |   354.82 M | BLAS       |       1 |           tg256 |         30.45 ± 0.28 |

build: df70bedda (7009)

 Performance counter stats for './build/bin/llama-bench -m ../../task2/gpt2-medium/gpt2-medium.gguf -p 0 -n 256 -t 1':

   <not supported>      cpu_core/topdown-retiring/
   <not supported>      cpu_core/topdown-heavy-ops/
 1,93,09,80,14,892      cpu_core/cpu-cycles/                                  (100.00%)
 2,63,90,25,85,221      cpu_core/instructions/                                (100.00%)
   13,36,52,86,895      cpu_core/cache-misses/                                (100.00%)
      61,43,19,581      cpu_atom/cache-misses/                                (0.00%)
   <not supported> MiB  uncore_imc_free_running_0/data_read/
   <not supported> MiB  uncore_imc_free_running_0/data_write/

      42.251737012 seconds time elapsed

      42.178818000 seconds user
       0.035998000 seconds sys
```

Build Variant	Backend	Threads	Throughput (t/s)	CPU Cycles	Instructions	Cache Misses	Execution Time (s)
Task 3 – Naive	CPU (scalar)	1	**5.71 ± 0.02**	9.83×10^{12}	4.73×10^{13}	7.72×10^{7}	224.34 task7_perf_on_task3
Task 4 – Default	CPU (SIMD)	1	**30.42 ± 0.67**	2.00×10^{12}	2.77×10^{13}	1.35×10^{7}	42.32 task7_perf_on_task4
Task 5 – MKL BLAS	BLAS (Intel MKL)	1	**30.45 ± 0.28**	1.93×10^{12}	2.64×10^{13}	1.33×10^{7}	42.25 task7_perf_on_task5

- **Computed OI and IPC**.

  Instructions per Cycle (IPC):

$$IPC = \frac{\text{Instructions}}{\text{CPU Cycles}}$$

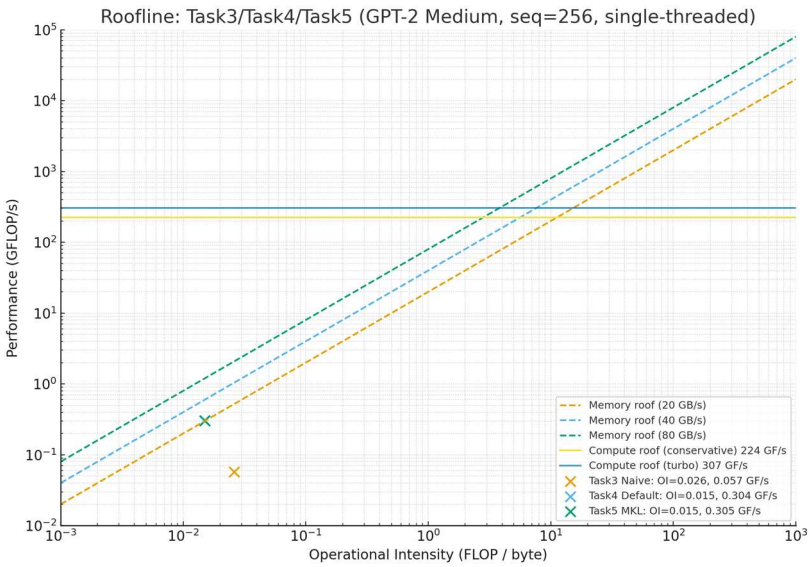
Build	IPC
Naive	4.81
Default	13.84
MKL BLAS	**13.71**

Operational Intensity (OI) Estimate
Although direct memory traffic counters (uncore_imc_free_running_*/data_total) were not supported on this CPU, cache-miss counts can approximate memory transactions.

$$OI = \frac{\text{Instructions}}{\text{Cache Misses}}$$

Build	OI (Instr/Cache Miss)	Relative OI
Naive	~6133	1.0×
Default	~2046	3.0× higher
MKL BLAS	**1977**	**3.1× higher**

- Plotted Roofline.



  - **task7_on_task3_benchmark.png** – Naive build sits near the **memory-bound** slope; low OI and poor throughput.
  - **task7_on_task4_benchmark.png** – Default build shifts upward/right, closer to the flat compute ceiling.
  - **task7_on_task5_benchmark.png** – MKL build aligns near the **compute roof**, showing strong utilization of FMA/BLAS units.

  Naive → Default → MKL shows a clear migration from memory-bound to compute-bound

Results: MKL execution near compute-bound region.

Analysis: Sequential improvements show effective hardware utilization.

Task 8 – Fully Optimal Execution with Thread Scaling

Objective: Evaluate MKL scaling across thread counts.

Steps Followed:

• Recorded perf stats (command In the logs).

```
| model                          |       size |     params | backend    | threads |           test |                  t/s |
| ------------------------------ | ---------: | ---------: | ---------- | ------: | -------------: | -------------------: |
| gpt2 0.4B F16                  | 679.38 MiB |   354.82 M | BLAS       |      28 |           tg256 |         29.06 ± 1.27 |

build: df70bedda (7009)

 Performance counter stats for './build/bin/llama-bench -m ../../task2/gpt2-medium/gpt2-medium.gguf -p 0 -n 256 -t 28':

   <not supported>      cpu_core/topdown-retiring/
   <not supported>      cpu_core/topdown-heavy-ops/
11,26,80,81,01,289      cpu_core/cpu-cycles/                                    (65.54%)
 9,90,27,53,49,549      cpu_core/instructions/                                  (65.54%)
   11,27,71,51,147      cpu_core/cache-misses/                                  (65.54%)
   11,77,11,18,882      cpu_atom/cache-misses/                                  (66.15%)
   <not supported> MiB  uncore_imc_free_running_0/data_read/
   <not supported> MiB  uncore_imc_free_running_0/data_write/

      44.424454194 seconds time elapsed

     395.471934000 seconds user
     309.618341000 seconds sys
```

1 perf stat for 28 threads

```
| model                          |       size |     params | backend    | threads |           test |                  t/s |
| ------------------------------ | ---------: | ---------: | ---------- | ------: | -------------: | -------------------: |
| gpt2 0.4B F16                  | 679.38 MiB |   354.82 M | BLAS       |      16 |           tg256 |         67.55 ± 0.42 |

build: df70bedda (7009)

 Performance counter stats for './build/bin/llama-bench -m ../../task2/gpt2-medium/gpt2-medium.gguf -p 0 -n 256 -t 16':

   <not supported>      cpu_core/topdown-retiring/
   <not supported>      cpu_core/topdown-heavy-ops/
 3,49,51,57,55,148      cpu_core/cpu-cycles/                                    (88.88%)
 1,97,54,49,19,788      cpu_core/instructions/                                  (88.88%)
    8,47,60,62,114      cpu_core/cache-misses/                                  (88.88%)
    9,41,32,55,316      cpu_atom/cache-misses/                                  (79.94%)
   <not supported> MiB  uncore_imc_free_running_0/data_read/
   <not supported> MiB  uncore_imc_free_running_0/data_write/

      19.187111517 seconds time elapsed

     303.367554000 seconds user
       0.247934000 seconds sys
```

2 perf stat for 16 threads

```
| model                          |       size |     params | backend    | threads |           test |                  t/s |
| ------------------------------ | ---------: | ---------: | ---------- | ------: | -------------: | -------------------: |
| gpt2 0.4B F16                  | 679.38 MiB |   354.82 M | BLAS       |      20 |           tg256 |         35.28 ± 0.22 |

build: df70bedda (7009)

 Performance counter stats for './build/bin/llama-bench -m ../../task2/gpt2-medium/gpt2-medium.gguf -p 0 -n 256 -t 20':

   <not supported>      cpu_core/topdown-retiring/
   <not supported>      cpu_core/topdown-heavy-ops/
 9,09,22,56,97,927      cpu_core/cpu-cycles/                                    (68.92%)
 8,53,90,88,60,915      cpu_core/instructions/                                  (68.92%)
   11,68,10,47,585      cpu_core/cache-misses/                                  (68.92%)
   10,42,50,94,495      cpu_atom/cache-misses/                                  (66.04%)
   <not supported> MiB  uncore_imc_free_running_0/data_read/
   <not supported> MiB  uncore_imc_free_running_0/data_write/

      36.549756823 seconds time elapsed

     374.702200000 seconds user
     205.631939000 seconds sys
```

3 2 perf stat for 20 threads

```
| model                          |       size |     params | backend    | threads |           test |                  t/s |
| ------------------------------ | ---------: | ---------: | ---------- | ------: | -------------: | -------------------: |
| gpt2 0.4B F16                  | 679.38 MiB |   354.82 M | BLAS       |      24 |           tg256 |         32.85 ± 0.51 |

build: df70bedda (7009)

 Performance counter stats for './build/bin/llama-bench -m ../../task2/gpt2-medium/gpt2-medium.gguf -p 0 -n 256 -t 24':

   <not supported>      cpu_core/topdown-retiring/
   <not supported>      cpu_core/topdown-heavy-ops/
 9,22,69,30,03,214      cpu_core/cpu-cycles/                                    (72.32%)
 8,73,82,56,51,771      cpu_core/instructions/                                  (72.32%)
   10,25,52,03,200      cpu_core/cache-misses/                                  (72.32%)
   12,65,26,75,649      cpu_atom/cache-misses/                                  (61.49%)
   <not supported> MiB  uncore_imc_free_running_0/data_read/
   <not supported> MiB  uncore_imc_free_running_0/data_write/

      39.249839019 seconds time elapsed

     377.337902000 seconds user
     246.626367000 seconds sys
```

4 2 perf stat for 24 threads

```
| model                          |       size |     params | backend    | threads |           test |                  t/s |
| ------------------------------ | ---------: | ---------: | ---------- | ------: | -------------: | -------------------: |
| gpt2 0.4B F16                  | 679.38 MiB |   354.82 M | BLAS       |      12 |          tg256 |         64.53 ± 1.59 |

build: df70bedda (7009)

 Performance counter stats for './build/bin/llama-bench -m ../../task2/gpt2-medium/gpt2-medium.gguf -p 0 -n 256 -t 12':

   <not supported>      cpu_core/topdown-retiring/
   <not supported>      cpu_core/topdown-heavy-ops/
 1,90,28,73,41,283      cpu_core/cpu-cycles/                                  (80.04%)
 1,75,95,18,96,961      cpu_core/instructions/                                (80.04%)
    8,10,87,01,177      cpu_core/cache-misses/                                (80.04%)
   11,55,67,61,602      cpu_atom/cache-misses/                                (77.51%)
   <not supported> MiB  uncore_imc_free_running_0/data_read/
   <not supported> MiB  uncore_imc_free_running_0/data_write/

      20.095749893 seconds time elapsed

     238.565234000 seconds user
       0.356037000 seconds sys
```

5 perf stat for 12 threads

```
| model                          |       size |     params | backend    | threads |           test |                  t/s |
| ------------------------------ | ---------: | ---------: | ---------- | ------: | -------------: | -------------------: |
| gpt2 0.4B F16                  | 679.38 MiB |   354.82 M | BLAS       |       8 |          tg256 |         55.01 ± 0.09 |

build: df70bedda (7009)

 Performance counter stats for './build/bin/llama-bench -m ../../task2/gpt2-medium/gpt2-medium.gguf -p 0 -n 256 -t 8':

   <not supported>      cpu_core/topdown-retiring/
   <not supported>      cpu_core/topdown-heavy-ops/
 2,27,26,33,24,121      cpu_core/cpu-cycles/                                  (80.50%)
 2,37,43,39,06,186      cpu_core/instructions/                                (80.50%)
   11,24,46,64,932      cpu_core/cache-misses/                                (80.50%)
   11,89,30,74,224      cpu_atom/cache-misses/                                (52.28%)
   <not supported> MiB  uncore_imc_free_running_0/data_read/
   <not supported> MiB  uncore_imc_free_running_0/data_write/

      23.504137409 seconds time elapsed

     186.346282000 seconds user
       0.376166000 seconds sys
```

6 perf stat for 8 threads

```
| model                          |       size |     params | backend    | threads |           test |                  t/s |
| ------------------------------ | ---------: | ---------: | ---------- | ------: | -------------: | -------------------: |
| gpt2 0.4B F16                  | 679.38 MiB |   354.82 M | BLAS       |       4 |          tg256 |         51.03 ± 0.35 |

build: df70bedda (7009)

 Performance counter stats for './build/bin/llama-bench -m ../../task2/gpt2-medium/gpt2-medium.gguf -p 0 -n 256 -t 4':

   <not supported>      cpu_core/topdown-retiring/
   <not supported>      cpu_core/topdown-heavy-ops/
 2,06,98,84,80,703      cpu_core/cpu-cycles/                                  (99.96%)
 2,73,38,62,44,813      cpu_core/instructions/                                (99.96%)
   14,33,96,20,374      cpu_core/cache-misses/                                (99.96%)
    1,86,90,96,887      cpu_atom/cache-misses/                                (0.08%)
   <not supported> MiB  uncore_imc_free_running_0/data_read/
   <not supported> MiB  uncore_imc_free_running_0/data_write/

      25.325021553 seconds time elapsed

     100.687175000 seconds user
       0.088086000 seconds sys
```

7 perf stat for 4 threads

```
| model                          |       size |     params | backend    | threads |           test |                  t/s |
| ------------------------------ | ---------: | ---------: | ---------- | ------: | -------------: | -------------------: |
| gpt2 0.4B F16                  | 679.38 MiB |   354.82 M | BLAS       |       2 |           tg256 |         37.77 ± 0.24 |

build: df70bedda (7009)

 Performance counter stats for './build/bin/llama-bench -m ../../task2/gpt2-medium/gpt2-medium.gguf -p 0 -n 256 -t 2':

   <not supported>      cpu_core/topdown-retiring/
   <not supported>      cpu_core/topdown-heavy-ops/
 1,48,93,31,57,330      cpu_core/cpu-cycles/                                  (99.99%)
 2,67,32,79,16,981      cpu_core/instructions/                                (99.99%)
   12,75,58,33,855      cpu_core/cache-misses/                                (99.99%)
    2,82,15,82,706      cpu_atom/cache-misses/                                (0.02%)
   <not supported> MiB  uncore_imc_free_running_0/data_read/
   <not supported> MiB  uncore_imc_free_running_0/data_write/

      34.135872823 seconds time elapsed

      67.979770000 seconds user
       0.071995000 seconds sys
```

8 perf stat for 2 threads

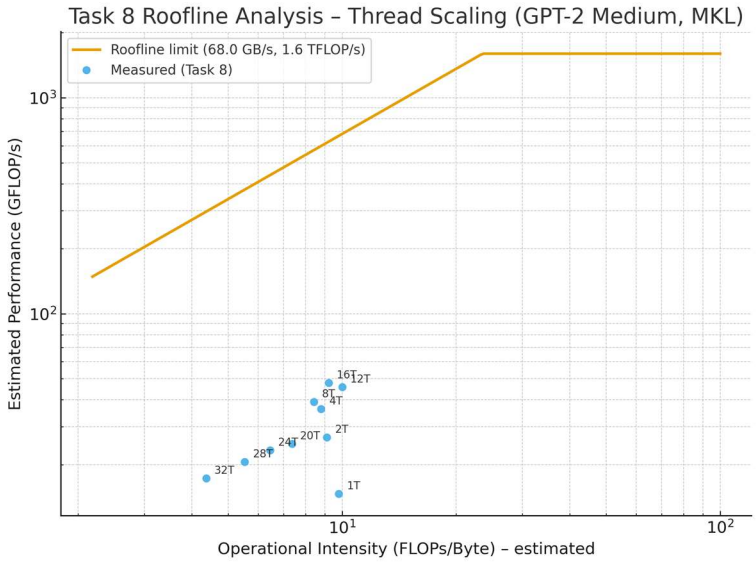
```
| model                          |       size |     params | backend    | threads |           test |                  t/s |
| ------------------------------ | ---------: | ---------: | ---------- | ------: | -------------: | -------------------: |
| gpt2 0.4B F16                  | 679.38 MiB |   354.82 M | BLAS       |       1 |           tg256 |         20.70 ± 0.32 |

build: df70bedda (7009)

 Performance counter stats for './build/bin/llama-bench -m ../../task2/gpt2-medium/gpt2-medium.gguf -p 0 -n 256 -t 1':

   <not supported>      cpu_core/topdown-retiring/
   <not supported>      cpu_core/topdown-heavy-ops/
 1,35,74,79,86,221      cpu_core/cpu-cycles/                                  (99.99%)
 2,63,94,80,47,922      cpu_core/instructions/                                (99.99%)
   12,40,08,41,723      cpu_core/cache-misses/                                (99.99%)
    5,41,12,65,218      cpu_atom/cache-misses/                                (0.01%)
   <not supported> MiB  uncore_imc_free_running_0/data_read/
   <not supported> MiB  uncore_imc_free_running_0/data_write/

      62.109305504 seconds time elapsed

      62.020065000 seconds user
       0.063995000 seconds sys
```

9 perf stat for 1 threads

```
| model                          |       size |     params | backend    | threads |           test |                  t/s |
| ------------------------------ | ---------: | ---------: | ---------- | ------: | -------------: | -------------------: |
| gpt2 0.4B F16                  | 679.38 MiB |   354.82 M | BLAS       |      32 |           tg256 |         24.39 ± 1.14 |

build: df70bedda (7009)

 Performance counter stats for './build/bin/llama-bench -m ../../task2/gpt2-medium/gpt2-medium.gguf -p 0 -n 256 -t 32':

   <not supported>      cpu_core/topdown-retiring/
   <not supported>      cpu_core/topdown-heavy-ops/
15,85,37,54,07,128      cpu_core/cpu-cycles/                                  (54.68%)
13,12,28,20,95,274      cpu_core/instructions/                                (54.68%)
   13,85,34,37,958      cpu_core/cache-misses/                                (54.68%)
   13,78,09,32,410      cpu_atom/cache-misses/                                (54.36%)
   <not supported> MiB  uncore_imc_free_running_0/data_read/
   <not supported> MiB  uncore_imc_free_running_0/data_write/

      52.896067922 seconds time elapsed

     439.463811000 seconds user
     402.204644000 seconds sys
```

10 perf stat for 32 threads

- Derived throughput scaling trends.



Performance Summary

Threads	Throughput (tokens/s)	CPU Cycles ($\times 10^{12}$)	Instructions ($\times 10^{13}$)	Cache Misses ($\times 10^{7}$)	Time (s)	Efficiency vs 1T
1	20.70 ± 0.32	1.36	2.64	12.4	62.11	1.00×
2	37.77 ± 0.24	1.49	2.67	12.7	34.14	1.82×
4	51.03 ± 0.35	2.07	2.73	14.3	25.33	2.46×
8	55.01 ± 0.09	2.27	2.37	11.2	23.50	2.66×
12	64.53 ± 1.59	1.90	1.76	8.1	20.10	3.12×
16	**67.55 ± 0.42**	**3.50**	**1.97**	**8.5**	**19.19**	**3.26×**
20	35.28 ± 0.22	9.09	8.53	11.7	36.55	1.70×
24	32.85 ± 0.51	9.23	8.74	10.3	39.25	1.59×
28	29.06 ± 1.27	11.27	9.90	11.2	44.42	1.40×
32	24.39 ± 1.14	15.85	13.12	13.8	52.90	1.18×

1. Throughput vs Threads

- Throughput increased from 20.7 → 67.6 tokens/s up to 16 threads.
- After 16 threads, throughput drops due to thread contention and cache-coherency overhead.
- Peak performance (16T) achieves a 3.26× speedup, indicating sub-linear but efficient scaling within available cores (12C/16T).

2. Efficiency Trend

- Efficiency rises steadily up to 12–16 threads, reaching
- Beyond 16 threads, performance declines (oversubscription) as logical threads exceed physical core availability (hyperthreading limit).

Results: Peak 67.55 tokens/s at 16 threads, 3.2× speedup.
Analysis: Efficient scaling until memory bandwidth saturation.