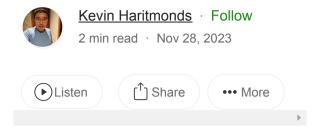
## How to run Llama 2 LLM in C



I was intrigued how does <u>ChatGPT</u> work. How can a computer generate a relevant, coherent answers to our question. I found an excellent presentation by Andrej Karpathy [1] in which he said Large Language Model (LLM) can just be two files. The parameters file (say Llama 2 Chat model from Meta AI) and a 500-lines of C code (<u>run.c</u> from same author [2]).

In this write up, I will describe how we can achieve this. Why Llama 2 model? Llama 2 is perhaps the most powerful open weight model today (2023). Why C? I wanted to learn how the text generation works, so a pure C code might be better as there is no library which may hide complexity.

- 1. Here I am using Linux Ubuntu-22.04 (from WSL 2).
- 2. Create our working directory:

```
$ mkdir -p ~/workspace/llm
$ cd ~/workspace/llm/
```

3. Download Meta llama-2-7b-chat model:

```
$ sudo apt update
$ sudo apt install -y git-lfs
$ git lfs clone https://huggingface.co/TheBloke/Llama-2-7B-Chat-fp16
```

This downloads around 27 GB of model files.

Warning: It is recommended to download the official model directly from <a href="https://ai.meta.com/llama/">https://ai.meta.com/llama/</a> and register. Then you must use "--meta-llama" argument instead of "--hf" on export.py step below.

4. Convert the model into .bin file:

```
$ wget https://raw.githubusercontent.com/karpathy/llama2.c/master/export.py
$ wget https://raw.githubusercontent.com/karpathy/llama2.c/master/model.py
$ sudo apt install -y python3-pip
$ pip3 install numpy torch transformers
$ python3 export.py llama-2-7b-chat.bin --hf ./Llama-2-7B-Chat-fp16
```

This operation requires 40 GB of RAM. It produces "llama-2-7b-chat.bin" file with 27 GB of size.

5. Convert the tokenizer.model into .bin file:

```
$ wget https://raw.githubusercontent.com/karpathy/llama2.c/master/tokenizer.py
$ pip3 install sentencepiece
$ python3 tokenizer.py --tokenizer-model=./Llama-2-7B-Chat-fp16/tokenizer.model
$ mv ./Llama-2-7B-Chat-fp16/tokenizer.bin .
```

It produces "tokenizer.bin" file.

6. Compile and execute run.c:

```
$ wget https://raw.githubusercontent.com/karpathy/llama2.c/master/run.c
$ gcc -Ofast run.c -lm -o run
$ ./run ./llama-2-7b-chat.bin -z ./tokenizer.bin -n 0 -m chat
```

7. (OPTIONAL) Speed up by utilizing multi CPU cores via OpenMP:

```
$ sudo apt install -y clang libomp-dev
$ clang -Ofast -fopenmp -march=native run.c -lm -o run
$ OMP_NUM_THREADS=8
$ ./run ./llama-2-7b-chat.bin -z ./tokenizer.bin -n 0 -m chat
```

Example conversations:

```
$ ./run ./llama-2-7b-chat.bin -z ./tokenizer.bin -n 0 -m chat
Enter system prompt (optional): Bill Gates
User: Who is his wife?
Assistant: Bill Gates is married to Melinda French Gates. They were married in 1994 and have }

<s>
User: When was the marriage?
Assistant: Bill Gates and Melinda French were married on January 1, 1994.

<s>
User: Where was it?
Assistant: Bill Gates and Melinda French were married in Hawaii, specifically on the island of the island on the island on the island on the island of the island on the island on the island of the island on the island of the island of
```

## Pretty cool, huh?

[1] Andrej Karpathy — [1hr Talk] Intro to Large Language Models