Llama 2

# Introduction

Llama2 is an open-source large language model developed by Meta AI.

<https://ai.meta.com/llama/>

<https://ai.meta.com/llama/get-started/>

Base Llama2 has three different sizes – 7B, 13B, and 70B

Llama-2-7b – 7 billion parameters

Llama-2-13b = 13 billion parameters

Llama-2-70b – 70 billion parameters

# Talk to Llama2

<https://www.llama2.ai/>

# GitHub Repository

[facebookresearch/llama: Inference code for LLaMA models (github.com)](https://github.com/facebookresearch/llama)

**Quantization**

<https://medium.com/@rakeshrajpurohit/model-quantization-with-hugging-face-transformers-and-bitsandbytes-integration-b4c9983e8996>

8bit

<https://arxiv.org/abs/2208.07339>

Memory Use

<https://huggingface.co/docs/transformers/main/en/model_doc/llama#transformers.LlamaForCausalLM>

https://huggingface.co/docs/transformers/main/en/model\_doc/llama2

Loading the model

from transformers import LlamaForCausalLM, LlamaTokenizer

tokenizer = LlamaTokenizer.from\_pretrained("/output/path")

model = LlamaForCausalLM.from\_pretrained("/output/path")

**CPU RAM to host the whole model in float16 precision or 2 bytes**

7B is then 7\*2 == 14 GB RAM

13B is then 13 \*2 == 26 GB RAM

70B is then 70 \*2 == 140 GB RAM

**To run the model in full precision on the GPU**

Full precision is float 32 or 4 bytes. 4 \* 7 == 28 GB GPU

13B is then 13 \*4 == 52 GB RAM

70B is then 70 \*4 == 280 GB RAM

Quantization to the rescue!

https://medium.com/@rakeshrajpurohit/model-quantization-with-hugging-face-transformers-and-bitsandbytes-integration-b4c9983e8996

Quantization is a technique used to reduce the precision of numerical values in a model. Instead of using high-precision data types, such as 32-bit floating-point numbers, quantization represents values using lower-precision data types, such as 8-bit integers. This process significantly reduces memory usage and can speed up model execution while maintaining acceptable accuracy.

13 b – **load\_using\_8bit=True**

7B is then 7\*1 == 7 GB RAM

13B is then 13 \*1 == 13 GB RAM

70B is then 70 \*1 == 70 GB RAM

# Hugging Face

<https://huggingface.co/meta-llama>

[moel]-hf – “converted for the Hugging Face Transformers format

[model]-chat is a fine tuned version of the corresponding base llama2.

A screenshot of a computer

Description automatically generated

# Coding

<https://github.com/facebookresearch/llama-recipes/blob/main/examples/Getting_to_know_Llama.ipynb>

1. Generative AI architectures
2. Chatbot architectures,
3. prompt engineering,
4. RAG (Retrieval Augmented Generation),
5. Fine-tuning
6. and more.

# **Quantize Llama**

<https://towardsdatascience.com/quantize-llama-models-with-ggml-and-llama-cpp-3612dfbcc172>

To allow LLMs to run on consumer hardware, a strategy called quantization has been developed by making the weight precision smaller. This increases the inference speed and uses less memory. [Georgi Gerganov](https://github.com/ggerganov/llama.cpp)

Many of quantization artifacts are hosted on hugging face by the author of a Llamacpp – a

[Llama.cpp | 🦜️🔗 Langchain](https://python.langchain.com/docs/integrations/llms/llamacpp)

[localmodels/Llama-2-7B-Chat-ggml · Hugging Face](https://huggingface.co/localmodels/Llama-2-7B-Chat-ggml)