LLM · Text Generation

How to run Meta's LLaMA on your computer (Windows, Linux tutorial)

By Tudor · March 30, 2023

Contents [hide] [1]

On March 3rd, user **Samsonov** leaked Meta's LLaMA model on 4chan's technology board (g) revealing access to it. It had seen advanced but not formal support in the Meta-official LLaMA GitHub repo.

This means LLaMA is the most powerful language model available to the public.

LLaMA specifications:

- There are four different pre-trained LLaMA models, with 7B (billion), 13B, 30B, and 65B parameters.
- Meta reports that the LLaMA 13B model outperforms GPT-3 in most benchmarks.
- Meta reports the 65B model is on par with Google's PaLM and DeepMind's Chinchilla.

4-bit LLaMA Installation

- 4-bit quantization is the method for running the size of models on they can run on less powerful hardware.
- Thanks to the efforts of many developers, **we can now run 4-bit LLaMA on most computer hardware.**

Here's some user-reported requirements for each model:

Model	Minimum Video RAM	Minimum Cpu RAM	Cpu cores recommended	4GB/Gbyte on disk*
LLaMA-7B	5.50GB	9GB	RTX 1660, 2060, AMD 5700 XT, RX 6600, 6700 XT, 6800, 6900 XT, 3060, 3060 Ti, 3070, 3070 Ti, 3080, 3090, 4070, 4080, 4090	10 GB
LLaMA-13B	9.5GB	18GB	AMD 6900xt, RTX 2060 12GB, 3060, 3060 12GB, 3080, 4070	20 GB
LLaMA-30B	19.5GB	40GB	RTX 3080 20GB, A4500, A5000, 3090, 4090, 6000, Tesla V100, P40	40 GB
LLaMA-65B	38.5GB	80GB	AMD MI200, 2x3090, 2x4090, A40, RTX A6000, 8000, Tesla Ton	128 GB

Windows & Linux Are Tier Minimal, please look at these instructions instead.

1. Install Prerequisites

Build Tools for Visual Studio 2019

Download "2019 Visual Studio and other products" (requires creating a Microsoft account). You must download the 2019 version.

In the Visual Studio Build Tools installer, check the Desktop development with C++ option and install.

Visual Studio Installer
Desktop development with C++
Installation details
Windows 10 SDK (10.0.19041.0)
C++ Clang tools for Windows (12.0.0 - x64/x86)
MSVC v142 - VS 2019 C++ x64/x86 build tools (v14.29)
C++ Clang-cl for v142 build tools (x64/x86)
C++ CMake tools for Windows
C++ ATL for v142 build tools (x86 & x64)
C++ MFC for v142 build tools (x86 & x64)
C++ Modules for v142 build tools (x64/x86 - experimental)
Testing tools core features - Build Tools
C++ AddressSanitizer
Live Share
IntelliCode
Python development
Node.js development
Azure Cloud Services (formerly Azure Tools)
Azure Data Lake and Stream Analytics (formerly Azure Tools)
Azure Stream Analytics (formerly Azure Tools)
Azure Cosmos DB (formerly Azure Tools)
Azure DevOps (formerly Visual Studio Team Services)
Azure Functions and Web Jobs (formerly Azure Tools)
Azure Data Lake (formerly Azure Tools)
Azure Blob Storage (formerly Azure Tools)
Azure Table and Queues (formerly Azure Tools)
Azure Service Fabric

I have the problem anybody knows why:
python.exe -u -m main --port 8188 --listen --lowvram CUDA not enabled
usage: server.py [-h] [--listen [IP]] [--port PORT] [--tls-keyfile TLSKEYFILE] [--tls-certfile TLSCERTFILE] [--enable-cors-header [ORIGIN]] [--max-upload-size MAXUPLOADSIZE] [--extra-model-paths-config PATH [PATH ...]] [--output-directory OUTPUTDIRECTORY] [--temp-directory TEMPDIRECTORY] [--input-directory INPUTDIRECTORY] [--auto-launch] [--disable-auto-launch] [--cuda-device DEVICE_ID] [--cuda-malloc] [--disable-cuda-malloc] [--dont-upcast-attention] [--force-fp32] [--force-fp16] [--bf16-unet] [--fp16-unet] [--fp8_e4m3fn-unet] [--fp8_e5m2-unet] [--fp16-vae] [--fp32-vae] [--bf16-vae] [--vae-in-cpu] [--vae-in-fp16] [--vae-in-bf16] [--vae-ref-tiling] [--vae-tiling] [--vae-ref-decoder] [--vae-decoder] [--vae-ref-encoder] [--vae-encoder] [--force-channels-last] [--directml [DIRECTML_DEVICE]] [--oneapi-device-selector SELECTOR_STRING] [--disable-ipex-optimize] [--preview-method [none,auto,latent2rgb,taesd]] [--preview-size PREVIEW_SIZE] [--cache-classic] [--cache-lru CACHE_LRU] [--cache-ram [CACHE_RAM]] [--fast [{fp16_accumulation,fp8_matrix_mult,cublas_ops}*]] [--mmap-torch-files] [--disable-mmap] [--dont-print-server] [--quick-test-for-ci] [--windows-standalone-build] [--disable-metadata] [--disable-assets-autoscan] [--enable-compress-response-body] [--comfy-api-base URL] [--database-url URL] [--enable-manager] [--disable-manager-ui] [--enable-manager-legacy-ui] [--disable-manager-legacy-ui] [--enable-manager-queue-ui] [--disable-manager-queue-ui] [--enable-manager-queue-workers N] [--disable-manager-queue-workers] [--enable-manager-queue-worker-ui] [--disable-manager-queue-worker-ui] [--enable-manager-queue-worker-ui-only] [--disable-manager-queue-worker-ui-only] [--enable-manager-queue-worker-ui-only-keep-alive N] [--disable-manager-queue-worker-ui-only-keep-alive] [--enable-manager-queue-worker-ui-only-keep-alive-minutes N] [--disable-manager-queue-worker-ui-only-keep-alive-minutes] [--write-log [PATH]] [--comfy-to-cloud-log-level {CRITICAL,ERROR,WARNING,INFO,DEBUG,NOTSET}] [--comfy-to-cloud-log-stdout] [--comfy-to-cloud-log-stderr] [--comfy-to-cloud-log-fmt FMT] [--comfy-to-cloud-log-datefmt DATEFMT] [--comfy-to-cloud-log-no-color] [--comfy-to-cloud-log-no-file] [--comfy-to-cloud-log-no-stdout] [--comfy-to-cloud-log-no-stderr] [--comfy-to-cloud-log-stdout-level {CRITICAL,ERROR,WARNING,INFO,DEBUG,NOTSET}] [--comfy-to-cloud-log-stderr-level {CRITICAL,ERROR,WARNING,INFO,DEBUG,NOTSET}] [--comfy-to-cloud-log-file-level {CRITICAL,ERROR,WARNING,INFO,DEBUG,NOTSET}] [--comfy-to-cloud-log-file-path PATH] [--comfy-to-cloud-log-file-max-bytes N] [--comfy-to-cloud-log-file-backup-count N] [--comfy-to-cloud-log-file-no-rotation] [--comfy-to-cloud-log-file-rotation-when {s,m,h,d,w0,w1,w2,w3,w4,w5,w6,midnight}] [--comfy-to-cloud-log-file-rotation-interval N] [--comfy-to-cloud-log-file-rotation-utc] [--comfy-to-cloud-log-file-no-rotation-utc] [--comfy-to-cloud-log-file-encoding ENCODING] [--comfy-to-cloud-log-file-no-encoding] [--comfy-to-cloud-log-file-mode {a,w,r,x,a+,w+,r+,x+}] [--comfy-to-cloud-log-file-no-mode] [--comfy-to-cloud-log-file-delay] [--comfy-to-cloud-log-file-no-delay] [--comfy-to-cloud-log-file-utf8] [--comfy-to-cloud-log-file-no-utf8] [--comfy-to-cloud-log-file-no-stdout-redirect] [--comfy-to-cloud-log-file-no-stderr-redirect] [--comfy-to-cloud-log-file-no-exceptions] [--comfy-to-cloud-log-file-no-traceback] [--comfy-to-cloud-log-file-no-traceback-limit] [--comfy-to-cloud-log-file-no-locals] [--comfy-to-cloud-log-file-no-locals-max-repr N] [--comfy-to-cloud-log-file-no-locals-max-depth N] [--comfy-to-cloud-log-file-no-locals-max-length N] [--comfy-to-cloud-log-file-no-locals-max-string N] [--comfy-to-cloud-log-file-no-locals-max-collection N] [--comfy-to-cloud-log-file-no-locals-max-dict N] [--comfy-to-cloud-log-file-no-locals-max-dataclass N] [--comfy-to-cloud-log-file-no-locals-max-namedtuple N] [--comfy-to-cloud-log-file-no-locals-max-tuple N] [--comfy-to-cloud-log-file-no-locals-max-set N] [--comfy-to-cloud-log-file-no-locals-max-frozenset N] [--comfy-to-cloud-log-file-no-locals-max-deque N] [--comfy-to-cloud-log-file-no-locals-max-array N] [--comfy-to-cloud-log-file-no-locals-max-mapping N] [--comfy-to-cloud-log-file-no-locals-max-sequence N] [--comfy-to-cloud-log-file-no-locals-max-iterable N] [--comfy-to-cloud-log-file-no-locals-max-attrs N] [--comfy-to-cloud-log-file-no-locals-max-depth-reached N] [--comfy-to-cloud-log-file-no-locals-max-recursion N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-min N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-max N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-min N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-max-repr N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-max-repr-length N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-max-repr-length-max N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-max-repr-length-min N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-max-repr-length-max-repr N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-max-repr-length-max-repr-length N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-max-repr-length-max-repr-length-max N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-max-repr-length-max-repr-length-min N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-min N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-min N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-min N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-min N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-min N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-min N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-min N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-min N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-min N] [--comfy-to-cloud-log-file-no-locals-max-recursion-depth-reached-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr-length-max-repr