# RUN LLM Mistral from scratch ROCm

## Requirments

* AMD Mi50/MI100 32Gb VRAM
* Workstation 40 GB RAM, 200GB SSD, 750W Power supply
* Ubuntu 24.04 LTS HWE Kernel
* Install python 3.11 or 3.12

## Steps

### Get the most popular LLM Mistral

git lfs install  
git clone https://huggingface.co/mistralai/Mistral-7B-v0.1 mistral

### Preapre python environment for ROCm:

python3 -m venv .venv\_llm\_mistral  
source ./.venv\_llm\_mistral/bin/activate  
python -m pip install --upgrade pip  
pip install torch torchvision torchaudio --index-url https://download.pytorch.org/whl/rocm6.0  
pip install transformers accelerate  
python .\test\_rocm\_mistral.py

### Create script test\_rocm\_mistral.py:

from transformers import AutoTokenizer, AutoModelForCausalLM, pipeline   
import torch   
   
print("GPU available:", torch.cuda.is\_available())   
print("GPU name:", torch.cuda.get\_device\_name(0))   
   
tokenizer = AutoTokenizer.from\_pretrained("/home/sysadmin/llm/mistral")   
model = AutoModelForCausalLM.from\_pretrained(   
 "/home/sysadmin/llm/mistral",   
 torch\_dtype=torch.bfloat16   
).to("cuda")   
   
generator = pipeline(   
 "text-generation",   
 model=model,   
 tokenizer=tokenizer,   
 device=0 # Use GPU   
)   
   
print(generator("What you know about sun?", max\_new\_tokens=60)[0]["generated\_text"])

### Enjoy the result!