
LOADING GPTQ 4-BIT MODEL WITH EXLLAMA
HOW TO LOAD A 7B PARAMETERS MODEL IN < 4GB GPU VRAM
HTTPS://GITHUB.COM/INSIGHTBUILDER

STEPS TO LOAD THE EXLLAMA

- DOWNLOAD THE PYTHON CONFIG FILES FROM EXLLAMA REPO
- DOWNLOAD THE LORA & NEKO 4-BIT LORA MODEL, TOKENIZER AND GENERATOR
- DOWNLOAD LORA MODEL CONFIG FILE FROM STANFORD ALPACA REPOSITORY
- SEND THE CONFIG & MODELS THROUGH THE EXLLAMA CLASSES AS SHOWN IN THE FILE
- TRY GENERATING THE OUTPUT WITH LARGE CONTEXT WINDOW

BENEFITS

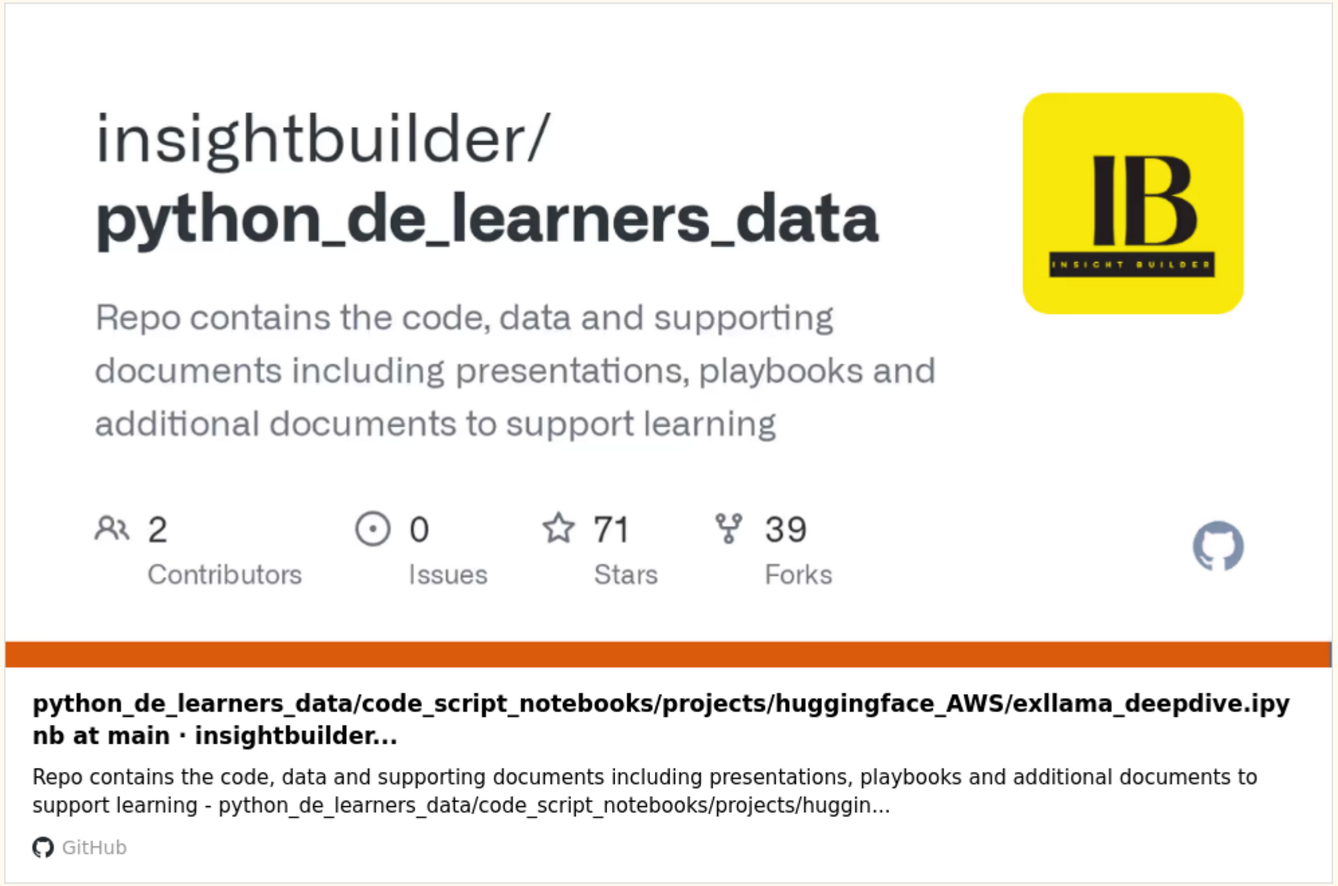
- INFERENCE IN A SINGLE T4 GPU PROVIDED FOR FREE IN GOOGLE COLAB / KAGGLE CAN BE USED FOR INFERENCE

7billion X 32 bits = 28 GB
7billion X 16bits = 14 GB
7 billion X 4 bit ~ 4 GB

- BUILDING CHATBOTS WITH GRADIO WITH LONGER MEMORY.
- CONNECTING WITH VECTOR QUERY WITH LARGER CONTEXT LENGTH FOR EASIER DOCUMENT ANALYSIS
- LESS TIME SPENT ON WAITING FOR THE INFERENCE & GET HANDS ON LEARNING ABOUT THE LLM

HTTPS://GITHUB.COM/INSIGHTBUILDER

CODE WITH EXPLANATION ON COLAB



HTTPS://GITHUB.COM/INSIGHTBUILDER

THANKS FOR WATCHING

REMEMBER TO PRACTICE WITH EXAMPLES

HTTPS://GITHUB.COM/INSIGHTBUILDER