

# LOADING GPTQ 4-BIT MODEL WITH EXLLAMA



**HOW TO LOAD A 7B  
PARAMETERS MODEL  
IN < 4GB GPU VRAM**

[HTTPS://GITHUB.COM/INSIGHTBUILDER](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**  
 $7\text{billion} \times 32 \text{ bits} = 28 \text{ GB}$   
 $7\text{billion} \times 16\text{bits} = 14 \text{ GB}$   
 $7 \text{ billion} \times 4 \text{ bit} \sim 4 \text{ GB}$
- **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**

# CODE WITH EXPLANATION ON COLAB

## insightbuilder/ python\_de\_learners\_data



Repo contains the code, data and supporting documents including presentations, playbooks and additional documents to support learning



2

Contributors



0

Issues



71

Stars



39

Forks



**python\_de\_learners\_data/code\_script\_notebooks/projects/huggingface\_AWS/exllama\_deepdive.ipynb at main · insightbuilder...**

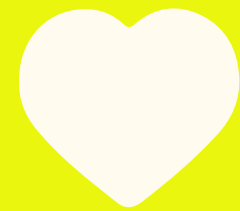
Repo contains the code, data and supporting documents including presentations, playbooks and additional documents to support learning - python\_de\_learners\_data/code\_script\_notebooks/projects/huggin...



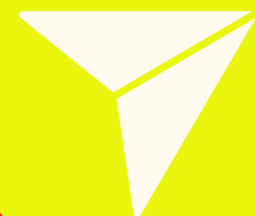
GitHub

# **THANKS FOR WATCHING**

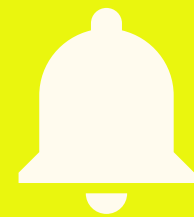
**REMEMBER TO PRACTICE WITH EXAMPLES**



**LIKE**



**SHARE**



**SUBSCRIBE**