**Things to Fix / Update**

1. **use\_auth\_token=True is deprecated**  
   Hugging Face no longer uses use\_auth\_token. Instead, you should log in with your **Hugging Face access token** (from https://huggingface.co/settings/tokens).
2. from huggingface\_hub import login
3. login("your\_hf\_token\_here")

Then remove use\_auth\_token=True from your from\_pretrained.

1. **Quantization (load\_in\_8bit) needs bitsandbytes**  
   If you use load\_in\_8bit=True, you must install bitsandbytes:
2. !pip install bitsandbytes accelerate

If you don’t install it, you’ll get import errors.  
On Colab Free (T4 GPU), this helps reduce memory usage.

1. **Full Working Colab Code**
2. !pip install transformers accelerate bitsandbytes huggingface\_hub
3. from huggingface\_hub import login
4. login("your\_hf\_token\_here") # replace with your Hugging Face token
5. from transformers import AutoTokenizer, AutoModelForCausalLM
6. model\_name = "meta-llama/Meta-Llama-3-8B-Instruct"
7. tokenizer = AutoTokenizer.from\_pretrained(model\_name)
8. model = AutoModelForCausalLM.from\_pretrained(
9. model\_name,
10. device\_map="auto", # lets HF automatically place on GPU
11. load\_in\_8bit=True # saves VRAM
12. )
13. def get\_llm\_response(prompt):
14. inputs = tokenizer(prompt, return\_tensors="pt").to(model.device)
15. outputs = model.generate(\*\*inputs, max\_new\_tokens=200)
16. return tokenizer.decode(outputs[0], skip\_special\_tokens=True)
17. print(get\_llm\_response("What is the capital of France?"))

⚠️ **Important Notes for Colab:**

* The **8B model is ~15GB** → even with 8bit, it’s heavy. On **Colab Free**, it may OOM (out of memory).
* Use **Colab Pro** (A100 GPU) for smoother runs.
* If it crashes, try a **smaller variant** (e.g., mistralai/Mistral-7B-Instruct-v0.2), which runs more easily in free Colab.

👉 Do you want me to also show you how to **swap this code to run a smaller model (like Mistral-7B)** so it won’t crash on free Colab GPUs?