

```
import gradio as gr
import torch
from transformers import AutoTokenizer


# Load model and tokenizer
model_name = "ibm-granite/granite-

tokenizer = AutoTokenizer.from_pre
model = AutoModelForCausalLM.from_
    model_name,
    torch_dtype=torch.float16 if t
    device_map="auto" # Automatic
)

# Function to get response
def ask_question(prompt):
    inputs = tokenizer(prompt, ret
    outputs = model.generate(
        **inputs,
        max_new_tokens=200,
        temperature=0.7,
        top_p=0.9,
        do_sample=True
    )
    return tokenizer.decode(output

# Gradio UI
demo = gr.Interface(
    fn=ask_question,
    inputs=gr.Textbox(lines=3, pla
    outputs="text",
    title="Edu Tutor AI",
    description="Personalized Lear

demo.launch()
```



```
/usr/local/lib/python3.12/dist-pac
The secret `HF_TOKEN` does not exi
To authenticate with the Hugging F
You will be able to reuse this sec
Please note that authentication is
warnings.warn(

tokenizer_config.json: 8.88k/? [00:00<00:0
vocab.json: 777k/? [00:00<00:00, 5.31ME
merges.txt: 442k/? [00:00<00:00, 2.74MB
tokenizer.json: 3.48M/? [00:00<00:00, 16
added_tokens.json: 100%
special_tokens_map.json: 100%
config.json: 100%
model.safetensors.index.json: 29.8k/? [00
Fetching 2 files: 0%
model-00002-of-
00002.safetensors: 100%
model-00001-of-
00002.safetensors: 34%
Loading checkpoint shards: 0%|
generation_config.json: 0%|
It looks like you are running Grad

Colab notebook detected. To show e
* Running on public URL: https://e

This share link expires in 1 week.
```



No interface is