```python
import gradio as gr
import torch
from transformers import AutoTokenizer, AutoModelForCausalLM

# Load model and tokenizer
model_name = "ibm-granite/granite-3.2-2b-instruct"

tokenizer = AutoTokenizer.from_pretrained(model_name)
model = AutoModelForCausalLM.from_pretrained(
    model_name,
    torch_dtype=torch.float16 if torch.cuda.is_available() else torch.float32,
    device_map="auto"  # Automatically places model on GPU/CPU
)

# Function to get response
def ask_question(prompt):
    inputs = tokenizer(prompt, return_tensors="pt").to(model.device)
    outputs = model.generate(
        **inputs,
        max_new_tokens=200,
        temperature=0.7,
        top_p=0.9,
        do_sample=True
    )
    return tokenizer.decode(outputs[0], skip_special_tokens=True)

# Gradio UI
demo = gr.Interface(
    fn=ask_question,
    inputs=gr.Textbox(lines=3, placeholder="Ask Edu Tutor AI..."),
    outputs="text",
    title="Edu Tutor AI",
    description="Personalized Learning with Generative AI + LMS Integration"
)

demo.launch()
```

```
/usr/local/lib/python3.12/dist-packages/huggingface_hub/utils/_auth.py:94: Use
The secret `HF_TOKEN` does not exist in your Colab secrets.
To authenticate with the Hugging Face Hub, create a token in your settings tab
You will be able to reuse this secret in all of your notebooks.
Please note that authentication is recommended but still optional to access pul
  warnings.warn(
tokenizer_config.json: 8.88k/? [00:00<00:00, 158kB/s]
vocab.json: 777k/? [00:00<00:00, 5.31MB/s]
merges.txt: 442k/? [00:00<00:00, 2.74MB/s]
tokenizer.json: 3.48M/? [00:00<00:00, 16.0MB/s]
added_tokens.json: 100% 87.0/87.0 [00:00<00:00, 797B/s]
special_tokens_map.json: 100% 701/701 [00:00<00:00, 6.12kB/s]
config.json: 100% 786/786 [00:00<00:00, 18.5kB/s]
model.safetensors.index.json: 29.8k/? [00:00<00:00, 1.85MB/s]
Fetching 2 files: 0% 0/2 [00:00<?, ?it/s]
model-00002-of-00002.safetensors: 100% 67.1M/67.1M [00:06<00:00, 9.61MB/s]
model-00001-of-00002.safetensors: 34% 1.68G/5.00G [00:42<01:52, 29.4MB/s]
Loading checkpoint shards: 0%| | 0/2 [00:00<?, ?it/s]
generation_config.json: 0%| | 0.00/137 [00:00<?, ?B/s]
It looks like you are running Gradio on a hosted Jupyter notebook, which requi

Colab notebook detected. To show errors in colab notebook, set debug=True in la
* Running on public URL: https://e500e8cf6d334d9a40.gradio.live

This share link expires in 1 week. For free permanent hosting and GPU upgrades
```

gradio

No interface is running right now