```
!pip install gradio torch transformers
```

```
Requirement already satisfied: pyyaml<7.0,>=5.0 in /usr/local/lib/python3.
Requirement already satisfied: ruff>=0.9.3 in /usr/local/lib/python3.12/di
Requirement already satisfied: safehttpx<0.2.0,>=0.1.6 in /usr/local/lib/p
Requirement already satisfied: semantic-version~=2.0 in /usr/local/lib/pyt
Requirement already satisfied: starlette<1.0,>=0.40.0 in /usr/local/lib/py
Requirement already satisfied: tomlkit<0.14.0,>=0.12.0 in /usr/local/lib/p
Requirement already satisfied: typer<1.0,>=0.12 in /usr/local/lib/python3.
Requirement already satisfied: typing-extensions~=4.0 in /usr/local/lib/py
Requirement already satisfied: uvicorn>=0.14.0 in /usr/local/lib/python3.1
Requirement already satisfied: fsspec in /usr/local/lib/python3.12/dist-pa
Requirement already satisfied: websockets<16.0,>=10.0 in /usr/local/lib/py
Requirement already satisfied: filelock in /usr/local/lib/python3.12/dist-
Requirement already satisfied: setuptools in /usr/local/lib/python3.12/dis
Requirement already satisfied: sympy>=1.13.3 in /usr/local/lib/python3.12/
Requirement already satisfied: networkx in /usr/local/lib/python3.12/dist-
Requirement already satisfied: nvidia-cuda-nvrtc-cu12==12.6.77 in /usr/loc
Requirement already satisfied: nvidia-cuda-runtime-cu12==12.6.77 in /usr/l
Requirement already satisfied: nvidia-cuda-cupti-cu12==12.6.80 in /usr/loc
Requirement already satisfied: nvidia-cudnn-cu12==9.10.2.21 in /usr/local/
Requirement already satisfied: nvidia-cublas-cu12==12.6.4.1 in /usr/local/
Requirement already satisfied: nvidia-cufft-cu12==11.3.0.4 in /usr/local/l
Requirement already satisfied: nvidia-curand-cu12==10.3.7.77 in /usr/local
Requirement already satisfied: nvidia-cusolver-cu12==11.7.1.2 in /usr/loca
Requirement already satisfied: nvidia-cusparse-cu12==12.5.4.2 in /usr/loca
Requirement already satisfied: nvidia-cusparselt-cu12==0.7.1 in /usr/local
Requirement already satisfied: nvidia-nccl-cu12==2.27.3 in /usr/local/lib/
Requirement already satisfied: nvidia-nvtx-cu12==12.6.77 in /usr/local/lib
Requirement already satisfied: nvidia-nvjitlink-cu12==12.6.85 in /usr/loca
Requirement already satisfied: nvidia-cufile-cu12==1.11.1.6 in /usr/local/
Requirement already satisfied: triton==3.4.0 in /usr/local/lib/python3.12/
```

```
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.12/dist-
Requirement already satisfied: markdown-it-py>=2.2.0 in /usr/local/lib/pyt
Requirement already satisfied: pygments<3.0.0,>=2.13.0 in /usr/local/lib/p
Requirement already satisfied: mdurl~=0.1 in /usr/local/lib/python3.12/dis
```

```
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
    device_map="auto" if torch.cuda.is_available() else None
)

# Fix pad token issue
if tokenizer.pad_token is None:
    tokenizer.pad_token = tokenizer.eos_token

# Function to generate response
def generate_response(prompt, max_length=512):
    inputs = tokenizer(prompt, return_tensors="pt", truncation=True, r

    if torch.cuda.is_available():
        inputs = {k: v.to(model.device) for k, v in inputs.items()}

    with torch.no_grad():
        outputs = model.generate(
            **inputs,
            max_length=max_length,
            temperature=0.7,
            do_sample=True,
            pad_token_id=tokenizer.eos_token_id
        )

    response = tokenizer.decode(outputs[0], skip_special_tokens=True)
    response = response.replace(prompt, "").strip()
    return response

# Function for explaining concepts
def concept_explanation(concept):
    prompt = f"Explain the concept of {concept} in detail with exampl
    return generate_response(prompt, max_length=800)

# Function for generating quiz
def quiz_generator(concept):
```

```
def quiz_generator(concept):
    prompt = f"Generate 5 quiz questions about {concept} with differe
    return generate_response(prompt, max_length=1200)

# Create Gradio interface
with gr.Blocks() as app:
    gr.Markdown("# 🎓 Educational AI Assistant")

    with gr.Tabs():
        # Concept Explanation tab
        with gr.TabItem("Concept Explanation"):
            concept_input = gr.Textbox(label="Enter a concept", placel
            explain_btn = gr.Button("Explain")
            explanation_output = gr.Textbox(label="Explanation", lines

            explain_btn.click(concept_explanation, inputs=concept_inpu

        # Quiz Generator tab
        with gr.TabItem("Quiz Generator"):
            quiz_input = gr.Textbox(label="Enter a topic", placeholde
            quiz_btn = gr.Button("Generate Quiz")
            quiz_output = gr.Textbox(label="Quiz Questions & Answers"

            quiz_btn.click(quiz_generator, inputs=quiz_input, outputs=

# Launch app with share link
app.launch(share=True)
```

```
/usr/local/lib/python3.12/dist-packages/huggingface_hub/utils/_auth.py:94: U
The secret `HF_TOKEN` does not exist in your Colab secrets.
To authenticate with the Hugging Face Hub, create a token in your settings t
You will be able to reuse this secret in all of your notebooks.
Please note that authentication is recommended but still optional to access
  warnings.warn(
```

tokenizer_config.json: 8.88k/? [00:00<00:00, 438kB/s]

vocab.json: 777k/? [00:00<00:00, 12.5MB/s]

merges.txt: 442k/? [00:00<00:00, 9.42MB/s]

tokenizer.json: 3.48M/? [00:00<00:00, 41.7MB/s]

added_tokens.json: 100% 87.0/87.0 [00:00<00:00, 3.29kB/s]

special_tokens_map.json: 100% 701/701 [00:00<00:00, 29.1kB/s]

config.json: 100% 786/786 [00:00<00:00, 20.4kB/s]

model.safetensors.index.json: 29.8k/? [00:00<00:00, 2.74MB/s]

Fetching 2 files: 100% 2/2 [01:33<00:00, 93.16s/it]

model-00002-of-00002.safetensors: 100% 67.1M/67.1M [00:03<00:00, 18.4MB/s]

model-00001-of-00002.safetensors: 100% 5.00G/5.00G [01:32<00:00, 51.7MB/s]

Loading checkpoint shards: 100% 2/2 [00:21<00:00, 8.73s/it]

generation_config.json: 100% 137/137 [00:00<00:00, 13.9kB/s]

```
Colab notebook detected. To show errors in colab notebook, set debug=True ir
* Running on public URL: https://a7dae4ee786275e2dc.gradio.live

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

No interface is running right now