```
!pip install transformers torch --quiet
```

```
from transformers import AutoModelForCausalLM, AutoTokenizer
import torch

# Load pre-trained DialoGPT
tokenizer = AutoTokenizer.from_pretrained("microsoft/DialoGPT-medium")
model = AutoModelForCausalLM.from_pretrained("microsoft/DialoGPT-medium")

# Initialize
chat_history_ids = None
step = 0

print("ChatBot is ready. Type your message. Type 'exit' to quit.\n")

# Continuous chat loop
while True:
    user_input = input("You: ")
    if user_input.lower() in ["exit", "quit", "bye"]:
        print("ChatBot: Goodbye! Have a great day!")
        break

    # Encode input
    new_input_ids = tokenizer.encode(user_input + tokenizer.eos_token, return_tensors="pt")

    # Append conversation history
    bot_input_ids = torch.cat([chat_history_ids, new_input_ids], dim=-1) if step > 0 else new_input_ids

    # Generate response
    chat_history_ids = model.generate(
        bot_input_ids,
        max_length=1000,
        pad_token_id=tokenizer.eos_token_id,
        do_sample=True,
        top_k=50,
        top_p=0.95,
        temperature=0.7
    )

    # Decode and print response
    response = tokenizer.decode(chat_history_ids[:, bot_input_ids.shape[-1]:][0], skip_special_tokens=True)
    print(f"ChatBot: {response}\n")

    step += 1
```

```
/usr/local/lib/python3.11/dist-packages/huggingface_hub/utils/_auth.py:94: UserWarning: 
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 (https://huggingface.co/settings/tokens), set it as secre
You will be able to reuse this secret in all of your notebooks.
Please note that authentication is recommended but still optional to access public models or datasets.
  warnings.warn(
tokenizer_config.json: 100%                                  614/614 [00:00<00:00, 19.8kB/s]
vocab.json: 100%                                       1.04M/1.04M [00:00<00:00, 7.68MB/s]
merges.txt: 100%                                       456k/456k [00:00<00:00, 20.4MB/s]
config.json: 100%                                      642/642 [00:00<00:00, 39.7kB/s]
pytorch_model.bin: 100%                                    863M/863M [00:05<00:00, 162MB/s]
model.safetensors: 100%                                    863M/863M [00:05<00:00, 222MB/s]
generation_config.json: 100%                                 124/124 [00:00<00:00, 6.31kB/s]
ChatBot is ready. Type your message. Type 'exit' to quit.

You: hello
The attention mask is not set and cannot be inferred from input because pad token is same as eos token. As a consequence, you may observ
ChatBot: Hey, I'm in the same situation.

You: you know i am mradul
ChatBot: Is that all?

You: yeah
ChatBot: Well I'm glad we can still talk.

You: i should not talk to you
ChatBot: I should not

You: yeah
ChatBot: Good talk

You: good night
ChatBot: good night

You: bye\
ChatBot: Good night
```