```
# Install required packages
# pip install transformers torch accelerate

from transformers import AutoTokenizer, AutoModelForCausalLM

# Load IBM Granite model from Hugging Face
model_id = "ibm-granite/granite-3.3-2b-instruct"
tokenizer = AutoTokenizer.from_pretrained(model_id)
model = AutoModelForCausalLM.from_pretrained(model_id)

# Input prompt
prompt = "What are the symptoms and treatment for malaria?"

# Tokenize the input
inputs = tokenizer(prompt, return_tensors="pt")

# Generate model response
outputs = model.generate(**inputs, max_new_tokens=100)

# Decode and print output
response = tokenizer.decode(outputs[0], skip_special_tokens=True)
print("\n💡 Response from IBM Granite model:")
print(response)
```