location

```
import gradio as gr
import time
from ctransformers import AutoModelforCasualLM

def load_llm():
  llm = AutoModelForCasualLM.from_pretrained(
    "codellama-13b-instruct.Q4_K_M.gguf",
    model_type= 'llama',
    max_new_tokens 1096,
    repetition_penality 1.13,
    temperature = 0.1,
  )
  return llm

def llm_function(message, chat_history):
  llm = load_llm()
  response = llm(
    message
  )
  output_texts = response
  return ooutput_texts


title = "Codellama 13B GGUF Demo"

examples = [
  "Write a python code to connect with a SQL database and list down all the tables.",
  "Write the python code to train a linear regression model using scikit learn.",
  "Write code to implement a binary tree implementation in C language.",
  "What are the benefits of the python programming language?"

]

gr.ChatInterface(
  fn = llm_function,
  title = title,
  examples = examples
)
```