LLM Input and Output based on prompt

Defining Model Name(LLaMA/Flan)

↓

Initializing model from transformer module

↓

Initializing Tokenizer(Auto) for the model selected

↓

Creating prompt

↓

Tokenizing the prompt and passing it to the model

↓

Generating tokens from the model output with an output limit

↓

Decoding the tokens through tokanizer and displaying it