

PYTHON FOR WEB

Introducing to Falcon 7b

Wiring Falcon 7B model to Django

Presented
By



A large, stylized green 'dj' logo, similar to the one in the main title, positioned in the top right corner of the image.

WHY?

- LEARN HOW THE LOADING LLM INTO A GPU WORKS IN THE BACKEND
- USE THE POST METHOD ON FORMS TO GET MORE INFORMATION FROM USERS
- STORE THE USER DATA & MODEL INFERENCE DATA IN THE DATABASE

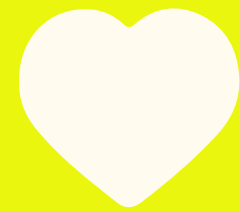
HOW?

- BUILD A PAGE WITH FORM THAT CAN SEND A POST REQUEST TO THE SERVER
- RESEARCH THE TRANSFORMERS MODEL LOADING PROCESS USING BITSNBYTES AND QUANTISATION(
- WRITE A VIEWS FUNCTION THAT ACCEPTS THE FORM DATA, STORES THEM IN DATABASE, PROCESS THE INFERENCE ON THE MODEL AND RETURNS THE RESULTS

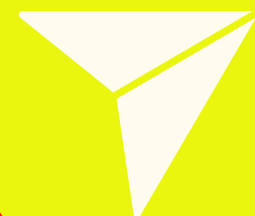
WHAT IT TAKES?

- TWO FUNCTION IN VIEWS.PY & ITS URLS PATH URLS.PY
- HTML BODY WITH A FORM CAPABLE OF POST REQUEST
- FALCON 7B REQUIRES GPU WITH 24GB, BUT QUANTISATION ALLOWS TO LOAD IN < 8GB

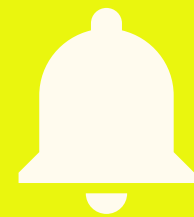
THANKS FOR WATCHING



LIKE



SHARE



SUBSCRIBE