HW 3

Create streaming chatbot that discusses a URL

Start from the Lab:

1. Copy lab3.py to HW3.py in the HWs folder. Add it to the HW app as a new page.

Define the options:

2. In the sidebar, provide an option for users to input two URLs.
3. In the sidebar, provide an option for users to pick the LLM to use (at least 3 vendors)
4. In the sidebar, provide an option for the type of conversation memory LLM to use (buffer of 5 questions, conversation summary, buffer of 5,000 tokens)
    a. Example user questions: Provide a summary of a sub topic, ask a specific fact, explain a concept, …

The user can select the LLM

5. For the sidebar menu that lets the user select which model to use:
    a. Have 2 from OpenAI (ex. 3.5 and 4o)
    b. Have 2 other LLMs (from 2 vendors that are not OpenAI or Microsoft)

2 URLs

6. Use both URL's (if 2 are provided), and the conversation history, to answer the question.
7. NOTE: The answer must be streamed to the user.

Deploy the app

8. Make sure to update requirements.txt as needed
9. Check to make sure your app (this new page) is working on the public URL

Questions for the HW

10. Evaluate the chat bot using two URLs on how to play cricket:
    - https://www.usatoday.com/story/graphics/2023/12/29/cricket-rules-scoring-explained/71570127007/
    - https://en.wikipedia.org/wiki/Cricket
11. Evaluate:
    a. Define 5 questions for your evaluation - explain why you selected the questions
    b. Explain which memory you think is best (and why)?
    c. Answer the questions using the following 6 different scenarios:
        i. just the first document, and both document
        ii. 3 different LLM models
    d. Compare the output generated for the 6 different scenarios
        i. Did having both documents improve the answers? Explain your answer
        ii. Which model was best? Explain your answer

What to submit:

- Submit the link and your short write-up providing your evaluation from above
- The HW3 python file (.py)