Ema Intern Take-Home Challenge

In this take-home challenge, you'll build a Natural Language Query Agent over some sample data. The purpose of this project is to demonstrate your ability to work with data, research techniques commonly used to solve such tasks and implement a trimmed down version of the approach to answer simple natural language queries.

Output: Upload a document with your approach and link to the github repository in the application form.

Here are the key details and requirements for the assessment:
You will utilize LLMs and open-source vector indexing and storage frameworks to answer simple questions over a small set of lecture notes and a table of LLM architectures. Although the core requirement is for generating simple, conversational answers over reference text, we encourage you to include methods or ideas for expanding the agent to handle more complex queries in your design (e.g. follow-up queries and conversational memory, citing references, consuming multiple classes worth of lecture notes). This will allow us to evaluate your understanding of the broader context and your ability to research and implement more advanced features.

Requirements and Guidelines:

1. Your system should take in a natural query and output an answer (conversational/abstractive) which sufficiently answers the question. For example, you can test with queries like:
    1. *"What are some milestone model architectures and papers in the last few years?"* → utilize text from introductory lecture note as well as the milestone papers table to put together an answer
    2. "*What are the layers in a transformer block?*"
    3. "*Tell me about datasets used to train LLMs and how they're cleaned*" (summarizing query)
2. Spend at most 4-5 hours on this project, and upon completion please push your work (README file, source code, dependencies, and any necessary instructions/design/notes to help us understand your implementation) to a github repository under your account and add Souvik (@souvik-sen) as a collaborator
3. Show your creativity in developing an intermediary representation: by this we mean how you're organizing raw data (web articles) into a structured textual/embedding-based format for storage and consumption by LLMs/your response synthesis layer. In your documentation, clearly state how you decided upon representing the raw lecture notes (organization of text/links/tables/etc).
4. Be as resourceful as possible – don't reinvent any wheels, feel free to use any open source libraries/frameworks as well as dev tools like CoPilot, some examples are langchain and llama-index

5. No need to train/fine-tune an architecture of your own, use existing frameworks/models/etc. Utilize LLMs/Multi-modal models by means of API calls, if you can't use your model of choice by API call, make note of it as future work.

Order of priorities:

1. A functional implementation is key – do not worry about implementing interfaces or setting up a deployment pipeline, but be sure to have some version of the challenge that can be demonstrated live in our follow-up call. Create a flexible intermediary representation, after which you can build a Q/A layer over it.
2. Clear and well documented description of your approach – again, creating a system architecture diagram isn't necessary, but have a clear description of the pipeline you've created, how you went about deciding on modules/frameworks/models to use, and what your approach's areas of improvement are.
3. Describe a deployment plan and scaling approaches (as # of lectures/papers grow, not # of users).

Bonuses:

- Citing references: display sections from lecture notes that were used to compose the final conversational answer. For example if the question is *"What are the layers in a transformer block"*, a citation may be a reference to the following excerpt: Citing references is about how the conversational answer was constructed to prove there wasn't any hallucination
- Conversational memory: supporting more than a single question (i.e. allowing for contextually aware follow-up questions)
- Generating summary of conversation session: after a series of questions and answers, putting together the equivalent of flashcards or study tips

Please let us know if you have any questions regarding the assessment or need further clarification on any aspect of the project. Souvik Sen, our co-founder is here to assist you