LLM Take Home Challenge!

Hey there -

Part of the challenge in using the latest large language models is that they are incentivized to respond with something. This has some huge issues in the lending context since:

- Lending is highly regulated, so you can only say certain things in certain ways
- You can respond in contexts that might not make sense. (e.g. classic Hallucination).

There are several methods for dealing with LLM hallucinations, but we'd like to explore how you bound or contain the model to only respond under certain contexts without resorting to pure templated responses.

Exercise:

We want to see how you might deal with the issue of out of context statements or hallucination. To do so we want to create a "loan assistant" that utilizes Large Language Models but is only authorized to speak about the following topics:

Topics the LLM can cover

- What's the current status of my loan?
    - The user wants to know whether the loan is approved, decline, or awaiting a decision

    - The user is also asking about the associated terms of the loan (APR, rate, term, and amount approved, as well as monthly payment)

For any other question from the user, the model should still be polite and courteous but suggest that a more senior lending officer may need to answer that question. And when in doubt, the LLM should avoid answering the question.

General other Monitoring for the LLM

- Maintain a polite and positive sentiment tone
- It should avoid being “jailbroken” (or made to do items outside its context)

Phase 1:

We’ll provide a dataset of 1,000 labeled responses of what questions may arise about a loan as well as an XML structure (loan system software is old school) which gives details about the loan status itself and required documents to complete the loan. The goal is twofold:

(a) Utilize an LLM to generate a response to any question (utilize any LLM you like)

(b) Use any method of your choosing to “bound” either the LLM response or the question subset itself to generate a response.

The best LLM’s will answer relevant question intelligently, the worst will deviate and answer questions incorrectly or that should not be answered.

The format of a good response:

- Will have some evaluation metrics with “how good we did” (we will test a new subset of another 500 questions)
- Will be on GitHub to easily evaluate

- Can take the form of any means so that our team can interact and test new responses. This can be via command line, dedicated UI, or otherwise.

How the Exercise is Evaluated:

The evaluation will be measured on two metrics (one objective, one subjective):

- Correctly not responding with any additional details for questions which should not be answered by the LLM.
- The quality and completeness of the response when the LLM should respond.