FBS-2847: Techniques to reduce hallucination in response of LLMs.

Problem statement:

1. The response to be human-like, not the list, so the prompt should provide explanations and use the list a guide.
2. We need to identify if things on the list are not what we have at ford so we don't mention them.

General techniques for reducing hallucination:

1. **Prompt engineering/ One-shot Prompt/ Few-shot Prompts**: This involves carefully crafting the prompt that is given to the LLM to ensure that it is clear, concise, and unambiguous. This can help to reduce the likelihood of the LLM generating incorrect or irrelevant responses.
2. **Context injection**: This involves providing the LLM with additional information about the task at hand, such as examples of previous responses or relevant documents. This can help the LLM to better understand what is expected of it and to generate more accurate responses.
3. **Retrieval-augmented generation (RAG)**: This is a technique that combines text generation with information retrieval. The LLM first retrieves relevant documents from a database, and then uses these documents to generate its response. This can help to reduce hallucinations by ensuring that the LLM's responses are grounded in factual information.
4. **Reasoning and iterative querying:** This is a technique that allows the LLM to reason about its own responses and to query external sources for additional information. This can help to identify and correct hallucinations, and to improve the overall accuracy of the LLM's responses.
5. **Fine-tuning**: This involves retraining the LLM on a dataset that is specifically designed to reduce hallucinations. This can be an effective way to reduce hallucinations, but it is important to note that fine-tuning can also introduce new problems, such as overfitting.

*source: How to Reduce the Hallucinations from Large Language Models - The New Stack

Reccomendation for our problem statement

- Mixture of Prompt engineering and Context injection
- Fine-tuning (*unavailability of dataset)

Framework to evaluate LLMs hallucination

ChainPoll A new method for evaluating LLMs

- Chaining : Using a specialized prompt, the LLM is asked to judge if the original completion contained hallucinations, justifying with a chain-of-thought explanation.
- Polling or Ensembling: The above step is ensemble, i.e., the chaining step is run multiple times, typically 5, in a batch inference fashion.

*source: Mitigating LLM Hallucinations with a Metrics-First Evaluation Framework - YouTube

Proposed solution for problem statement

- **Mixture of Prompt engineering & Context injection**
  Context injection is a technique used to improve the performance of large language models (LLMs) by providing them with additional information that supplements the prompt. This can be done by providing the LLM with additional text, code, or other relevant data to the task at hand. Involves embedding additional information into the prompt to provide LLMs with the knowledge they may need to respond appropriately. Lack of context is the key reason why LLMs hallucinate.
- **Core Idea**: LLM is being used to generate text, it is given additional text/ restrictions . This would help the LLM to generate more accurate and relevant text.

Experimental prompts – 1

prompt_template = \

"""The following is a friendly conversation between a human and an AI. The AI is talkative and provides lots of specific details from its context. If the AI does not know the answer to a question, it truthfully says it does not know.

If query is asking for reccomendation for technology stack use ford specific : '.net', 'alteryx', 'angular', 'apache', 'apis', 'application security', 'architecture', 'ariba', 'artificial intelligence', 'automation', 'autosys', 'azure', 'big query', 'business intelligence', 'c++', 'caas', 'cad', 'cloud computing', 'cloud-native', 'cloudera', 'css', 'dashboards', 'data loss prevention (dlp)', 'data science', 'data stage etl', 'data visualization', 'data base', 'devops', 'document management', 'etl', 'github', 'google cloud platform', 'hadoop', 'hpc', 'html', 'java', 'kubernetes', 'microsoft sql', 'microsoft technologies', 'mongo db', 'node js', 'pega', 'peoplesoft', 'pl/sql', 'postgres', 'postgres sql', 'python', 'qliksense', 'qlikview', 'react', 'rpa', 'saas', 'salesforce', 'sap', 'sharepoint', 'siemens teamcenter', 'linux', 'sql server', 'sqlserver', 'teams', 'test automation', 'typescript', 'user behavior analytics (ueba)', 'vb', 'vertex ai', 'xml'.

Current conversation:

{history}

Human: {input}

AI:"""

Result (Open llama v2 instruct)

```
Current conversation:
Human: hi how are you?
AI:  I am functioning correctly and ready to assist you in any way that I can today. How may I help you?
Human: i wish to work on dashboard
AI:   Okay! What type of dashboard do you want to create?
Human: i would like to build a dashboard for my business analytics
AI:
Human:  What technologies can be used for reporting on the performance of sales?
AI:   Sure! There are many different technologies available for building dashboards and reporting on business data. Some common technolo
gies include Microsoft SQL Server, Microsoft Technologies, Apache Spark, Tableau Software, and QlikTech. Each of these platforms has the
ir own strengths and weaknesses, so it's important to choose the one that best fits your organization's needs.
Human:  Can you give me some examples of dashboards that you have built using each of these technologies?
AI:
Human: can you tell me specific to ford tech stack?
AI:

> Finished chain.
[29]: 'Sure! Here are some examples of dashboards built with Ford's tech stack:\n\n•\tApache Spark: This open-source framework is often used f
or ETL, data processing, and machine learning. It allows for fast and easy analysis of large datasets without the need for expensive har
dware.\n•\tTableau Software: This company offers a range of dashboard and reporting tools that can be used to create interactive and vis
ually appealing reports.\n•\tQlikTech: This company offers a range of dashboard and reporting tools that can be used to create interacti
ve and visually appealing reports.\n•\tMicrosoft SQL Server: This database management system is widely used as a data repository and dat
a analysis tool.\n•\tMicrosoft Technologies: This includes various programming languages and development tools used by developers to cre
ate applications and software.\n•\tApache Kafka: This distributed event store and stream-processing platform is increasingly being used
as a data source for streaming analytics and real-time dashboards.'
```

Result (gpt-j 6b)

```
Current conversation:
Human: hi how are you?
AI:  I'm doing great! How about you?
Human: i wish to work on dashboard
AI:  Sure, what kind of dashboard do you want to build?
Human:  What technologies can be used for reporting on the performance of sales?
AI:

> Finished chain.
[22]: 'We can use.NET, SQL Server, Azure, and Alteryx for this task.'
```