Update for Guardrails Implementation for GenAI

1. **For Week (12-15 Sept):**
    a. **Text Quality Guardrail:**
        i. Exploring Langkit's "texstat" module for calculating metrics readability, complexity and grade level for determining text quality.
            1. This week, our team delved into the Langkit library's "texstat" module, which provides tools for assessing text quality. We are evaluating its suitability for calculating metrics related to readability, complexity, and grade level. These metrics are crucial for ensuring that the generated text meets appropriate quality standards.
        ii. Exploring how NLP task related metrics like ROGUE (Text Summarization), BELU (NLQ), Perplexity and others can be used for determining text quality when used as a guardrail for LLMs.
            1. We are also investigating the utilization of NLP task-specific metrics such as ROGUE for text summarization, BELU for Natural Language Questions (NLQ), and perplexity for language modeling. We aim to determine how these metrics can serve as effective guardrails to assess and ensure text quality in LLMs.
        iii. Exploring SummaC and QAFactEval for scenario-based text quality evalution.
            1. As part of our text quality assessment, we are exploring SummaC and QAFactEval, which are tools designed for scenario-based text quality evaluation. These tools will assist us in evaluating how well the generated text aligns with specific scenarios and facts.

    b. **Hallucination Guardrail:**
        i. Exploring FLARE framework paper for hallucination detection.
            1. In our effort to implement the hallucination guardrail, we are closely studying the FLARE framework paper, which provides insights and methodologies for detecting hallucination in text generated by language models.
        ii. Exploring RAG for hallucination detection
            1. We are also investigating the use of the RAG (Retrieval-Augmented Generation) framework to detect hallucination in the generated text. RAG integrates retrieval-based methods into the generation process, which can help identify discrepancies and hallucinations.
        iii. Exploring Bert-score, MQAG Framework and N-gram to detect hallucination in LLMs.
            1. Our exploration includes the evaluation of various techniques, including Bert-score, the MQAG (Multi-Question Answering Generation) Framework, and N-grams, to detect instances of hallucination in the output of LLMs.
        iv. Exploring Rail Framework for hallucination detection.
            1. Additionally, we are exploring the Rail Framework, which is designed for hallucination detection. This framework incorporates advanced techniques to identify and mitigate hallucination issues in generated text.

c. **Privacy Guardrail:**
    i. Exploring Opaque Prompts through Langchain.
        1. Our privacy guardrail efforts involve investigating the use of opaque prompts through Langchain, a technology designed to enhance privacy in AI systems. We aim to leverage Langchain for secure and private interactions with the model.
    ii. Exploring DelD-GPT for de-identification of PII data.
        1. We are exploring DelD-GPT, a solution for de-identifying Personally Identifiable Information (PII) data in the model's responses, ensuring compliance with privacy regulations.
    iii. Going through a research paper that uses Differential Privacy concept.
        1. As part of our privacy research, we are reviewing a research paper that leverages the concept of Differential Privacy. This concept is critical for preserving privacy when working with sensitive data.
    iv. Exploration of NER approaches for identification of PII data in prompts. To be further used masking and de-masking of PII data.
        1. To enhance privacy, we are exploring Named Entity Recognition (NER) approaches to identify PII data in prompts. This identification will enable us to implement effective masking and de-masking strategies to protect sensitive information.

2. **Plan for Week (19-22 Sept):**
    a. **Text Quality Guardrail:**
        i. Further exploration of above-mentioned approaches.
            1. In the upcoming week, we will continue to explore and refine the approaches mentioned above for text quality assessment.

    b. **Hallucination Guardrail:**
        i. Further exploration of above-mentioned approaches.
            1. We will build on our investigations from the previous week, focusing on enhancing the hallucination guardrail.

    c. **Privacy Guardrail:**
        i. Further exploration of above-mentioned approaches.
            1. Our privacy guardrail efforts will continue with a deeper dive into the approaches we've identified to ensure robust privacy safeguards for GenAI.