Guidehouse

October 24, 2023

Rod Fontecilla, Ph.D.

Navigating the Digital Labyrinth: Safeguarding the Use of Large Language Models

Introduction

Large Language Models (LLMs) and Generative AI stand at the forefront of innovation, shaping the very fabric of our digital interactions.

The Intelligence Community (IC) sees the transformative potential of LLMs – envision agents rapidly summarizing crucial intelligence data, contextualizing multifaceted information, and making swift, informed decisions.

But with great power comes an imperative for responsible and safe use.

In this talk, we present the future where AI tools harmoniously integrate with human agents, and where technology is both innovative and safe.

Unique Application in the IC

Guidehouse agrees with most observers that LLMs and Generative AI (GenAI) will increase the efficiency and effectiveness of traditional data collection and analysis. We also believe there are several non-traditional areas ripe for intense research where GenAI could potentially be tested against IC requirements in a way that could tremendously advantage US national security:

- Multiple methods of GenAI concurrently used for consolidating and analyzing thousands of streams of different, multi-modals of intelligence simultaneously, such as distinct types of IMINT, SIGINT (e.g., ELINT and COMINT), MASINT, OSINT, HUMINT, and much more
- GenAI as a technique to try to prevent denial and deception techniques and to spot anomalies that suggest something is off-track
- Recognizing that adversaries are leveraging similar technology, we see significant potential in offensive operations. For example, leveraging LLMs within a US IC organization to run an information operations campaign to spoof and exploit vulnerabilities in an adversary's tool

Due to the extreme sensitivity of these applications, they also require supporting methodologies that can identify and detect errors and biases prior to being deployed.

Development of Novel LLM Probing Methodologies

1. Input Fuzzing: Randomly altering inputs to see how the LLM responds, revealing hidden biases or problematic behaviors.

2. Adversarial Testing: Deliberately crafting inputs to mislead or trick the LLM into generating wrong or inappropriate outputs.

3. Attention Visualization: Mapping out which parts of the input the LLM focuses on while producing an output. This can help in understanding decision-making patterns.

4. Reverse Engineering: Deciphering the LLM's learned patterns by examining its outputs over a diverse set of inputs.

5. Model Explainability: Using tools and techniques to make the decision-making processes of LLMs transparent and understandable.

Probing Methods

Input Fuzzing: uncover vulnerabilities or defects by providing the system with unexpected or random data (called “fuzz”) and then monitoring for anomalies like crashes, undocumented behaviors, or memory leaks.

- **Application to LLMs**:
  - **Fault Discovery**: By providing random or semi-random inputs to an LLM, fuzzing can help discover how the model might fail or behave unexpectedly.
  - **Security**: It can uncover potential security threats where specific inputs could be used maliciously to derive unintended outputs.
  - **Model Robustness**: Fuzzing can be essential in understanding the boundaries of what an LLM can handle, thereby aiding in refining the model.

Attention Visualization:

- **Definition**: In the context of deep learning models, particularly transformers, attention mechanisms allow the model to focus on specific parts of the input when generating an output. Attention visualization is the process of graphically representing how and where the model is placing its focus during its operations.
- **Application to LLMs**:
  - **Understanding Decision-making**: Visualizing attention weights can infer which parts of an input influenced the LLM's response the most.
  - **Detecting Biases**: If an LLM consistently places high attention weights on biased or irrelevant parts of the input, it might indicate an underlying bias in the model.
  - **Model Interpretability**: Attention visualization can make LLMs more interpretable and user-friendly.

Agents act as a sentinel, ensuring LLM outputs are accurate, ethical, and safe

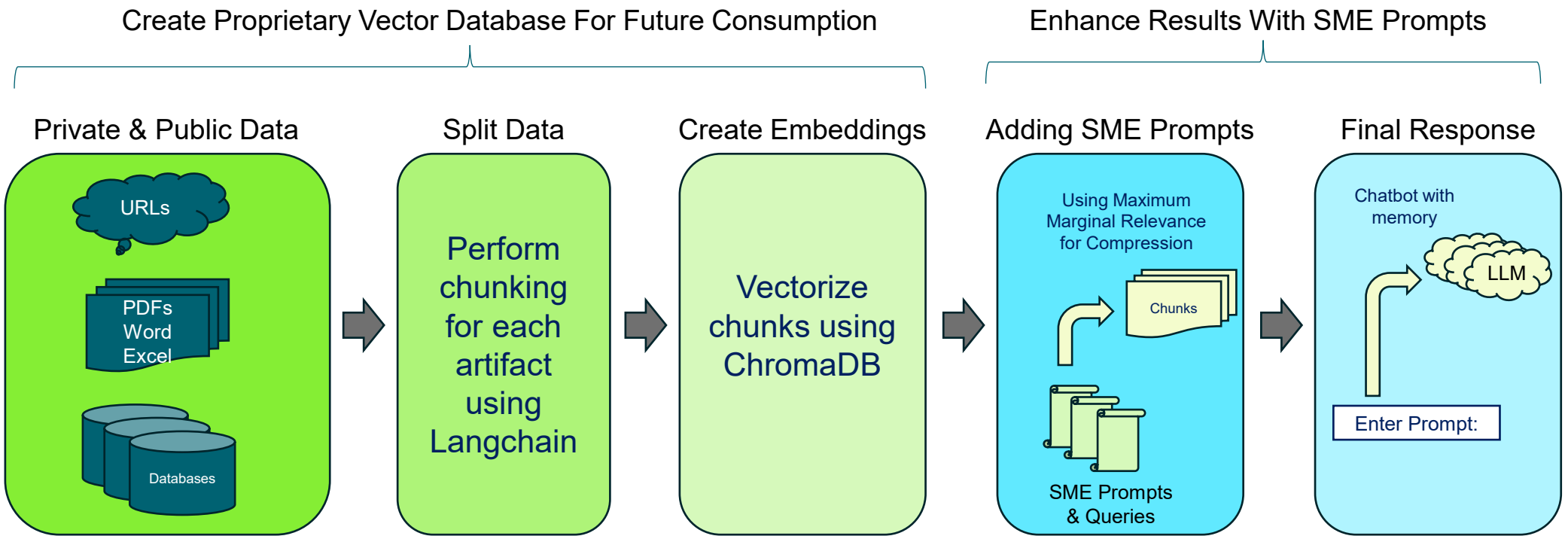
Real-time Monitoring and Intervention: Agents can be designed to monitor LLM outputs in real-time, evaluating them against predefined safety metrics or ethical benchmarks.

- **Application**:
  - **Instant Red Flags**: If the LLM produces a biased, incorrect, or potentially harmful output, the agent can instantly flag it.
  - **Auto-Correction**: Agents can either suggest corrections or automatically adjust the LLM's outputs before they are presented to the user.

Simulated Adversarial Testing: Agents can act as adversarial entities, deliberately trying to elicit erroneous or biased responses from the LLM.

- **Application**:
  - **Resilience Building**: By continuously challenging the LLM in this manner, it becomes more resilient to real-world adversarial attacks.
  - **Threat Landscape Mapping**: Understand potential vulnerabilities by observing how the LLM responds to these simulated attacks.

Retrieval Augmented Generation Methodology



Load Data into GH Azure Blob Storage

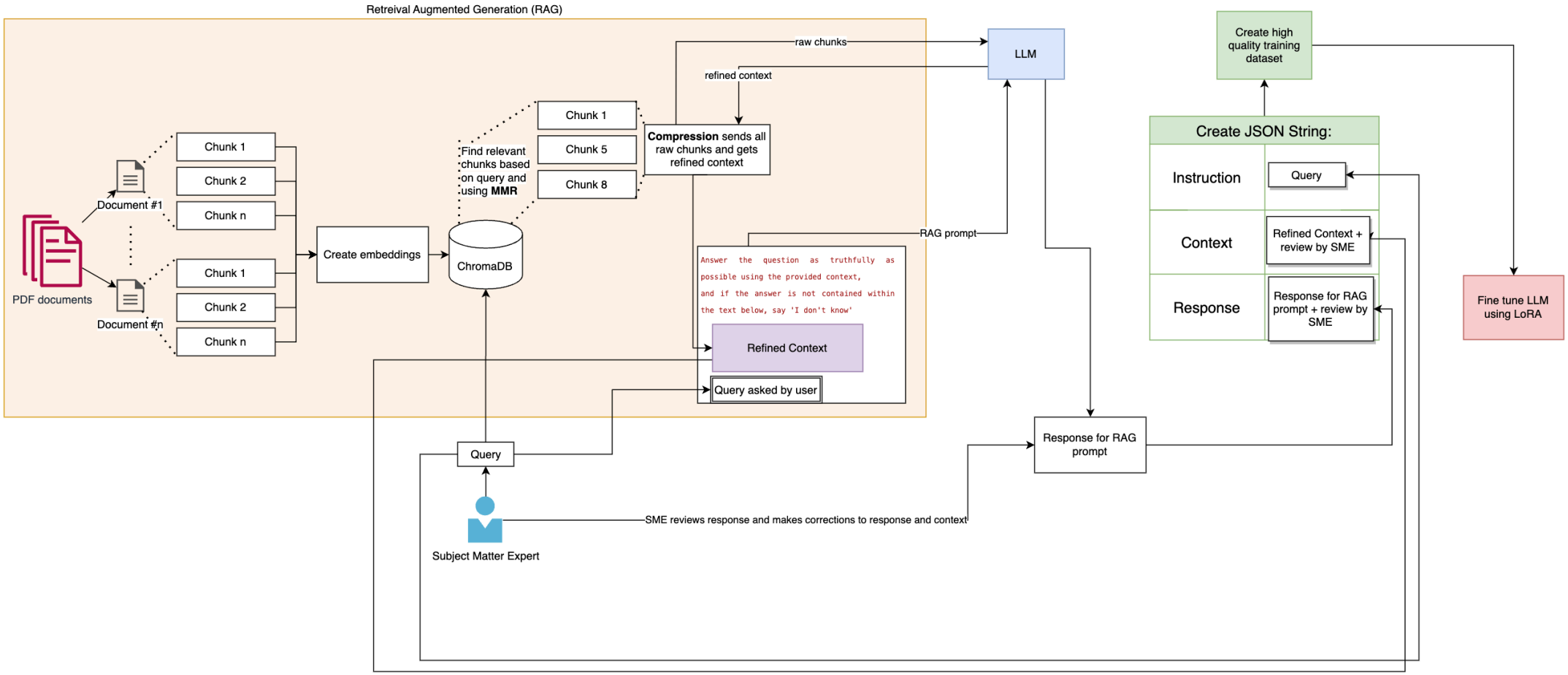
Use Recursive Character Text Splitter

Get Relevant Context for Model

Pass query and relevant chunks to get better response from LLM

Select LLM Based on Use Case

Fine-tuning LLM Methodology with Proprietary Data



Rod Fontecilla, Ph.D.
Partner, Chief Innovation Officer
rod.fontecilla@guidehouse.com
(240) 271-1682

Thank You