**Understanding RAG Systems**

A Retrieval-Augmented Generation (RAG) system combines the power of two components – a document retrieval system and a generative language model. The retrieval system first searches for relevant documents from a given collection, and then the language model generates answers based on these documents.

As we integrate a RAG system into our eDiscovery offering, it’s crucial to understand its capabilities and limitations. This memo aims to clarify where our RAG system is likely to succeed and where it may struggle, ensuring effective usage and setting realistic expectations.

**Strengths of a RAG System**

*Specific Factual Questions*

Questions that work well with RAG systems are those that can be answered by information **contained in the documents of the collection**. These include:

Specific factual questions: "What is the capital of France?"

Detailed inquiries that align with the collection's content: "What are the health benefits of green tea according to recent studies?"

**Limitations of a RAG System**

*Broad Summaries*

A RAG system cannot provide an accurate summary of the entire collection unless there is a specific document in the collection that summarizes it. So, asking for a general overview of all documents is not feasible.

*Quantitative Analysis of Documents*

Questions like "how many documents cover the topic of climate change?" are not ideal. **A RAG system isn't designed to quantify the number of documents related to a topic; it's built to find and use content from those documents to answer questions.**

*Highly Technical Queries*

If the collection lacks deep, specialized content on a subject, the system may not effectively answer complex queries in that domain.

*Creative or Abstract Responses*

The RAG system is not suited for tasks that require creativity or abstract thinking. Its responses are based on existing information in the document collection.

*Ambiguous Questions*

Vague queries can lead to unclear answers. For instance, asking "What should we focus on in the case?" is too open-ended for the system to provide a useful response.

**Hypothetical Question Assessments**

1. What happened in a certain location (geographic location)?  
     
   Assessment: Likely well-suited if the documents cover events or activities in that location. The system can retrieve information specific to the geographic area mentioned.
2. What was the type of business being conducted in a certain location (geographic location)?  
     
   Assessment: Well-suited if business activities in the specified location are documented. The system can identify and provide information on business types from the documents.
3. What was {project code name} within the company?  
     
   Assessment: Well-suited if there are documents describing the project. The system can identify information associated with the specific code name.
4. What are some examples of anticompetitive activity?  
     
   Assessment: Not well suited as the request is too general. The system can provide examples only if the activities are explicitly documented as being anticompetitive.
5. What are some effects of anticompetitive activity?  
     
   Assessment: Similar to the previous question, it's well-suited if the documents include discussions or analyses of these effects.
6. What are some examples of nefarious activity?  
     
   Assessment: May struggle due to the broad and subjective nature of "nefarious activity". Specificity in the question would help.
7. What events led to the incidents?  
     
   Assessment: Well-suited if the documents include a chronological account or analysis of the events leading to specific incidents.
8. Who were the witnesses, and what did they observe?  
     
   Assessment: Well-suited if witness statements or observations are documented as such.
9. Who was involved in a particular incident?  
     
   Assessment: Well-suited if the documents contain information about individuals involved in the incident.
10. Who has access to sensitive or relevant information?  
      
    Assessment: Well-suited if access details are documented.
11. Who else is involved in the incident or impacted by the incident?  
      
    Assessment: Well-suited if the documents contain this information.
12. Was [someone/one of the custodians] involved in the incident?  
      
    Assessment: Well-suited if involvement details of specific individuals are mentioned.
13. When did the incident occur?  
      
    Assessment: Well-suited if the documents include date and time details of the incident.
14. **How many** documents are related to the incident/project?  
      
    Assessment: Likely to struggle as RAG systems are not designed for quantitative analysis of document counts. They are built to answer factual questions about the contents of particular documents.
15. Provide a timeline of relevant activities to the incident?  
      
    Assessment: Well-suited only if a timeline or chronological order of events is documented.
16. Find **all** communication with an external entity.  
      
    Assessment: Not well suited.
17. Identify **all** the potentially privileged documents.  
      
    Assessment: Not well suited both because assessing privilege requires more information than is apparent on the face of a document, legal reasoning and, in this case, quantitative analysis of the corpus.
18. Find **all** documents that relate to an incident.  
      
    Assessment: Not well suited.
19. Find relevant documents related to an incident/event [within a date range].  
      
    Assessment: Likely not well suited. Requires legal reasoning.
20. Find **all** communication between [these custodians] [within a date range]  
      
    Assessment: Not well suited for a RAG system. A RAG system is built to answer factual questions where the answers are to be found in specific documents. It is not a general search tool.
21. Identify **all** documents which potentially relate to the topic of [\_\_\_\_\_]  
      
    Assessment: Not well suited.