## Evaluation and Analysis:

**The questions_answers.csv file in the directory is updated with the responses of my RAG pipeline. From the comparison of my pipeline's outputs to the reference responses, the following is observed:**

The Retrieval-Augmented Generation pipeline is generally functioning well, effectively retrieving both text blocks and tables in response to user queries. However, certain issues with table extraction and the disconnect between textual and tabular data have led to limitations in the agent's ability to accurately and comprehensively answer some queries. Specifically:

1. **Truncated or Distorted Tables:**
    - The current method of extracting tables using the unstructured library often results in tables being truncated, missing important rows, or becoming distorted. This incomplete or inaccurate tabular data significantly impacts the quality of retrieval and downstream responses.
    - **Impact:** Some queries are answered incorrectly due to missing or incomplete tabular data.
    - **Example:** A truncated table extracted as the context for one of the queries.

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | **Isokorb® Lange [mm]** | | |
| | **Bestückung bei** | | | **500** | | |
| | **Zugstabe/Druckstabe** | | | **2x6@12** | | |
| **Querkraftstabe** | **2x36** | **2x3 28** | **2x3@10** | **2x42 10** | **2x4 12** | |
| **Hin** | **bei CV35 [mm]** | **160** | **170** | **180** | **180** | **190** |
| **Hmin** | **bei CV50 [mm]** | **200** | **210** | **220** | **220** | **230** |

2. **Missing Link Between Text Blocks and Tables:**
    - Since tables are stored separately from the surrounding text, the context required to relate textual and tabular data is often missing. This is especially problematic as many tables lack explicit titles or headers, making it difficult for the retriever to associate them with the correct textual context.
    - **Impact:**
        - The agent is unable to answer some queries due to missing context.
        - When tables are retrieved without related text, the responses can be incomplete or inaccurate.

**Proposed Solutions**

**1. Improve Table Extraction Quality**

- **Problem:** Tables extracted using the unstructured library are often truncated or distorted.
- **Proposed Solution:**
    - Replace or complement the unstructured library with PyPDF for table extraction. PyPDF has demonstrated better performance in accurately extracting tables with complete rows and columns.

- Validate extracted tables against predefined completeness criteria before storing them in the docstore.

**2. Link Text Blocks and Tables**

- **Problem:** Text blocks and tables are stored separately, leading to a loss of context.
- **Proposed Solution:**
  - Bind text blocks and tables under the same section number during document ingestion. This ensures that they are stored and retrieved together.
  - Update the summarization step to include relevant surrounding text when summarizing tables. For instance:
    - Extract additional context from the same section as the table.
    - Concatenate the table summary with the surrounding text summary for coherent context.
  - During retrieval, fetch both the text and table summaries if either is identified as relevant to the query.

**3. Enhance Summarization for Better Retrieval**

- **Problem:** The summarization step does not effectively incorporate table context with surrounding text, further compounding retrieval challenges.
- **Proposed Solution:**
  - Use a fine-tuned summarization chain that:
    - Captures the relationship between tables and their surrounding text.
    - Outputs enriched summaries that integrate tabular data and contextual text.
  - Store these enriched summaries in the vectorstore for better semantic retrieval.

**Anticipated Benefits**

1. **Improved Data Quality:**
   - Accurate and complete table extraction ensures reliable data for retrieval and querying.
2. **Enhanced Retrieval Context:**
   - Linking text and tables under the same section improves the relevance of retrieved information.
   - Enriched summaries ensure that retrieved data is both comprehensive and contextual.
3. **Better Query Responses:**
   - With accurate and complete context, the agent is more likely to generate correct and comprehensive answers, reducing the number of unanswered or incorrect responses.