## Task 1

The default chunking strategy I used was recursive character text splitting. An additional strategy I would like to test out is character text splitting.

The primary motivation for these two chunking strategies is simplicity. When working in business, we typically develop a prototype before developing the final product. A prototype enables us to verify that the solution meets the business needs, communicates our vision of the solution to the business, helps mitigate risks, and saves on development costs. As such, given our provided use case, I would want to prototype a solution as quickly and efficiently as possible. Since much of our pre-existing code already uses recursive character text splitting, and as recursive character text splitting seems to perform well in comparison with other chunking methods, I used it for my default strategy.

Originally, I wanted to examine agentic chunking and/or semantic chunking as the alternative strategy. While these strategies were successfully executed, they suffered from complications. Both strategies were successfully tested. However, agentic chunking took an incredibly long time for our two documents, so it would not be a good strategy for a business with a high volume of documents. Both strategies suffered from issues that arose when trying to evaluate their performances with the RAGAS framework. Recent updates to Langchain make it incompatible with RAGAS. This is problematic as the code to execute semantic chunking requires the latest version of Langchain and therefore cannot execute successfully with RAGAS. Therefore, because of these complications, I chose character text splitting as the alternate strategy.

## Task 2

I went with a retrieval augmented pipeline using gpt-4o-mini as the model. The two reasons for the selection of these tools were time and cost. As the focus is developing a prototype, the business would want the cheapest option as soon as possible. Given that much of the code was already available and as gpt-4o-mini is a cheap, though efficient, option, I chose these tools to complete the task.

I attempted to build an agentic RAG system. I successfully built the graph, but I could not figure out how to get the graph to function properly with Chainlit. Perhaps with more time, I could have successfully implemented the graph, but because of the complications and difficulties I had, I defaulted to using the retrieval augmented pipeline.

Loom Video: [MidtermTask2 - a Hugging Face Space by acpotts - Google Chrome - 24 September 2024 (loom.com)](https://www.loom.com/share/dc94d9c22d5f41a7998dd51b2c1bb441)

### Task 3:

Table 1. Results from recursive chunking.

|  |  |  |  |
| --- | --- | --- | --- |
| Relevance | Faithfulness | Context Recall | Precision |
| 0.8297 | 0.7963 | 0.7458 | 0.7903 |

1. **Relevance (0.8297)**:
   * Relevance measures how relevant the generated responses are to the input queries. A score of 0.8297 suggests that the pipeline performs well at generating responses that are pertinent to the questions asked.
2. **Faithfulness (0.7963)**:
   * Faithfulness assesses how accurately the generated responses reflect the information in the source documents. A score of 0.7963 indicates that the pipeline maintains a high level of accuracy and truthfulness in its responses, though there is still some room for improvement.
3. **Context Recall (0.7458)**:
   * Context recall measures how well the pipeline retrieves and uses relevant context from the source documents. A score of 0.7458 shows that the pipeline is fairly effective at recalling and utilizing context, but there is potential to enhance this aspect further.
4. **Precision (0.7903)**:
   * Precision evaluates the correctness of the retrieved information. A score of 0.7903 indicates that the information retrieved by the pipeline is mostly accurate, but there might be some irrelevant or incorrect details included.

**Summary:**

Overall, the pipeline performs well across all metrics, with scores indicating strong relevance, faithfulness, and precision. However, there is some room for improvement, particularly in context recall and faithfulness, to ensure even higher accuracy and relevance in the generated responses.

## Task 4

Link: <https://huggingface.co/spaces/acpotts/AIE4midtermTask4>

As previously mentioned, the goal of a prototype is to quickly develop a working example so that the business can have a better understanding of the final product. Because the code already contained a working example with Snowflake arctic, I used that model for the embeddings. With more time, a proper comparison could be conducted with various embedding models.

## Task 5

Table 2. Chunking strategies comparison

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Strategy | Relevance | Faithfulness | Context Recall | Precision |
| Character | 0.8165 | 0.7735 | 0.8562 | 0.7559 |
| Recursive Character | 0.8297 | 0.7963 | 0.7458 | 0.7903 |

Based on the results from Table 2, there doesn’t seem to be a “bad” strategy. Recursive chunking performed a little better on 3 of the 4 metrics, but character chunking had a significantly higher recall score. Given the rigidity of character chunking, I would recommend recursive chunking, but either strategy would probably work for our use case.

## Task 6

**Story for the CEO to Tell at the Launch**

**Title: Embracing the Future with Ethical and Impactful AI**

**Introduction:** “Today marks a significant milestone in our journey towards harnessing the power of AI to drive real value for our company. With the board’s support, we have assembled a top-tier data science and engineering team to leverage the latest open-source technologies, including Retrieval-Augmented Generation (RAG) and intelligent agents.”

**The Vision:** “Our goal is clear: to achieve a substantial Return on Investment (ROI) from our research and development efforts. This isn’t just about innovation for innovation’s sake; it’s about creating tangible benefits for our business and stakeholders.”

**Addressing Concerns:** “We understand that the rapid advancement of AI brings with it concerns about ethics and utility. Our commitment is to build AI applications that are not only powerful but also ethical and beneficial for enterprises. We are dedicated to ensuring that our AI solutions are developed responsibly and transparently.”

**The Solution:** “One of the key initiatives we’ve undertaken is the development of a chatbot application. This tool will help our employees and stakeholders understand the evolving AI landscape, especially in the context of politics and governance. By allowing users to upload documents and ask questions, the chatbot will provide valuable insights and guidance.”

**Call to Action:** “Over the next month, we will be testing this application with 50 internal stakeholders. This will help us refine the tool and ensure it meets the needs of our users. We are excited about the potential of this application to empower our team and drive informed decision-making.”

**Conclusion:** “Together, we are not just keeping pace with the AI revolution; we are leading it. Let’s embrace this opportunity to innovate responsibly and create a brighter future for our company and society.”

**Incorporating Future Information**

To incorporate important updates like the 270-day update on the 2023 executive order on Safe, Secure, and Trustworthy AI, we can follow these steps:

1. **Regular Monitoring:**
   * Assign a team member to regularly monitor official sources such as the White House website and other government publications for updates on AI policies and guidelines.
2. **Automated Alerts:**
   * Set up automated alerts using tools like Google Alerts or RSS feeds to stay informed about new developments in AI regulations and policies.
3. **Content Integration:**
   * Develop a process to integrate relevant information into your AI models. This could involve updating the training data with new documents and guidelines, ensuring that the chatbot can provide up-to-date and accurate information.
4. **Periodic Reviews:**
   * Schedule periodic reviews of the chatbot’s content to ensure it remains current. This could be done quarterly or bi-annually, depending on the frequency of updates in the AI policy landscape.
5. **User Feedback:**
   * Encourage users to provide feedback on the chatbot’s responses. Use this feedback to identify gaps in the information and update the system accordingly.
6. **Collaboration with Legal and Compliance Teams:**
   * Work closely with your legal and compliance teams to ensure that all incorporated information aligns with regulatory requirements and company policies.