AI Solutions Architect: Capstone Project: Part 1

A large technology company wants to develop a state-of-the-art generative AI system capable of producing high-quality, coherent, and factual content on a wide range of topics on it's products. The system, which will replace the search capability on the corporate homepage, should leverage AI retrieval-based methods to gather relevant information from various sources and produce human-like outputs. The key requirements include:

1) Build a comprehensive knowledge base by retrieving and indexing information from trusted sources, such as product documents, technical documentation.

2) Implement advanced natural language processing techniques to understand user queries and extract relevant context and intent.

3) Utilize retrieval algorithms to identify and retrieve the most relevant information from the knowledge base based on the user's query.

4) Incorporate large language models and generation techniques to synthesize the retrieved information into coherent, fluent, and contextually appropriate responses.

5) Ensure factual accuracy of the output by cross-checking generated content against the knowledge bases.

6) Implement techniques to control the style, tone, and formality of the generated content based on the user's preferences or the context of the query.

7) Provide explainable and transparent outputs, clearly indicating the sources of information used.

8) Design a modular and extensible architecture that allows for easy integration of new data sources, language models, and generation techniques.

9) Implement robust security measures to prevent the system from generating harmful, biased, or offensive content.

10) Continuously monitor and evaluate the system's performance using human evaluations, automated metrics, and user feedback.

11) Ensure that the solution is only available to authorized users

12) Consider the costs for the solution. How do they scale with increases or decreases in use by end-users, increases in data store size. Ideally, costs will scale with use.

Your assignment is to consider the above, draft some designs for the solution and answer the following questions.

1. Create a conceptual design for your solution, showing it in the context of the customer's current system as you imagine it (based on the limited information that you have)

2. Create a high-level digram that shows the main processing-steps for the solution. Your starting point is the user query, the end point being to solution output back to the user. Ideally your diagram will indicate the flow of data through the processing steps. You choose the level of detail that want to use to explain your solution.

   Note: Not all aspects of the requirement need to be accounted for. The objective is to be able to explain to stakeholders how the main flow of solution works.

3. Briefly document 3-5 project specific risks and their mitigations. One sentence per risk, 1-3 sentences per mitigation.

4. Consider 1-3 notable tools (e.g., models types, frameworks, databases, inference tooling, etc.) that you might use for the solution. Then, using the high-level processing-steps diagram as reference, overlay the tools, to highlight your initial implementation thoughts.

5. You need to return to the customer with the above artifacts before you can ask questions of specifics, as they will be incommunicado. You will need to make assumptions, based on the above and your industry knowledge.

   Are there questions that you would like to ask, so as to better ground you design? You quite possibly have many. For the assignment, briefly note the 3 most important questions, that you would ask the customer, when you meet with them to help clarify your assumptions and/or concerns.

By all means use tooling such as ChatGPT, Anthropic Claude, and Amazon Q to help you with this task. To the extent that you do, review the answers generated such that you feel reasonably confident that they are valid.

Submit your answers in a single PDF document.