GenAI Implementation Technical Considerations

1) Need to walkthrough the Python scripts with Mohammed and Ben (Planned for 14/08/2024)

2) Try to find a .NET implementation for the Langchain and embedding aspect:
   https://github.com/jeastham1993/langchain-dotnet
   https://github.com/tryAGI/LangChain

3) Architectural Considerations:

   - Embeddings and Storage:
     Every time a corpus of documents is uploaded, we will create two sets of embeddings: one embedding for the multiple documents together that will be used to answer the questions in the current instance and the other set of embeddings will be individual embeddings for each document. The embeddings can be stored as '.pkl' files and these files will be stored in a blob storage in a structured folder structure to resemble the assessment organisation in the frontend to keep the relevant embeddings segregated. The embedding should only be stored if the user selects 'Store Documents' option, otherwise the embedding will be created on the fly and stored in memory and removed after the session ends for confidentiality purposes.

   - Prompt Libraries:
     We need a standard set of prompts for the various use cases that we cater for in Assessment Hub - need to create a table that we can fetch the relevant prompt from and insert the question to then feed to the chat model to generate an answer. Example:

     "You need to act as an assistant to help PwC Assessment Hub team answer the questions based on only the context provided to you. Be as detailed as possible in your answer. \
          Answer ONLY with the facts listed in the list of sources below, and only use the sources which you think are relevant to the question. \
          If there isn't enough information below, say you don't know. You must not generate answers that don't use the sources below. \
          You must ALWAYS include file name and the page numbers as your source for each fact you use in the response. \
          Use square brakets to reference the source, e.g. [Source 1 fileName.txt]. \
          Don't combine sources, list each source separately, e.g. [Source 0 fileName.doc][Source 3 fileName.pdf].\

          Question: {question}
          Context: {context}"

4) Use of GenAI SDK provided by PwC GIF shared service: PwC recommended to use for GenAI tools as they are looking to improve this suite of tools.

   Limitations:
   - Limited to text inputs only (no photos, audios)

- Only Python implementation provided for now