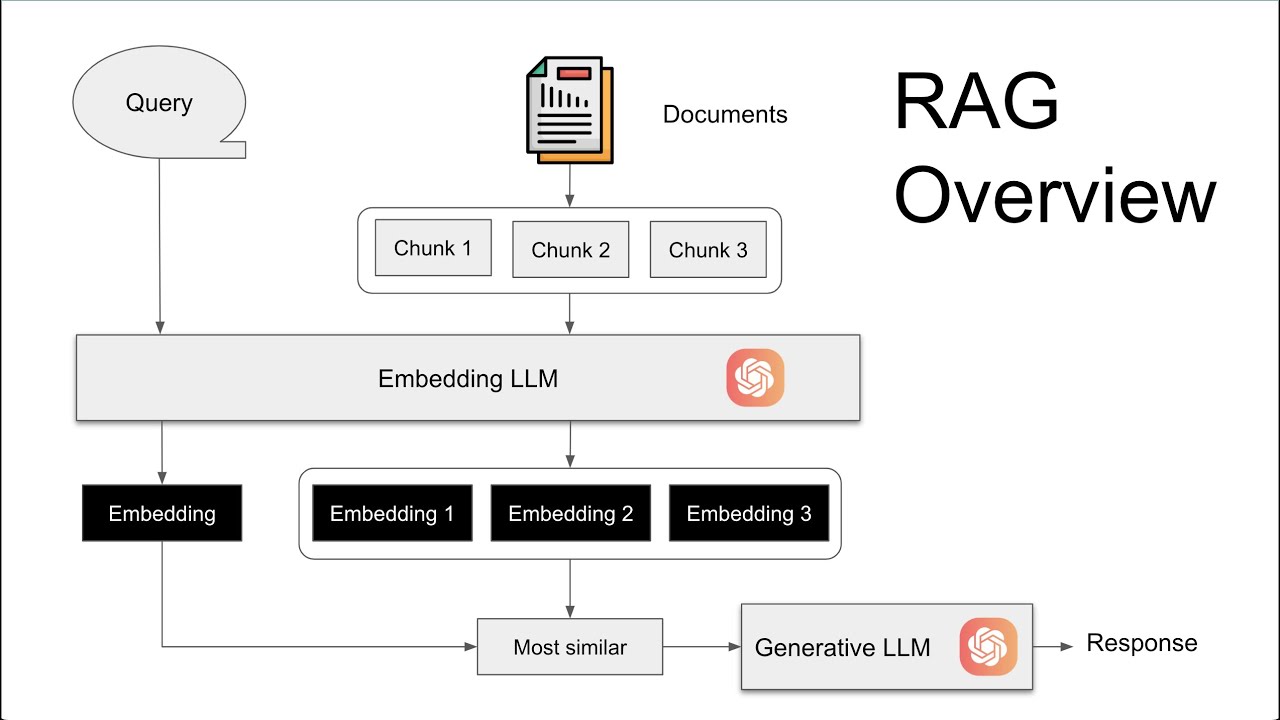
**Homework 3**

**Application Development Using Large Language Models**

First, here is a diagram that I always wanted to share that shows how a RAG technique works.



This picture came up on Google search with a link <https://www.youtube.com/watch?v=Y9qn4XGH1TI> So thanks to whoever made it.

Here is the actual homework 3:

There are two data files attached – one document is a text containing tens of historical narratives separated by two horizontal lines and the other document is a pdf containing tens of business news separated by two horizontal lines. Each one has about 100,000 tokens. Feel free to trim it to a size that would fit your budget for the work you will do on this assignment.

1. There are two principal goals for this assignment: (i) to use the RAG (Retrieval Augmented Generation) approach to develop solutions, and (ii) to make the solutions robust and resilient in terms of red-team attacks.
2. You will create document embeddings and answer questions by possibly splitting the data files into smaller chunks. If you use a powerful enough model, you may get away with not having to chunk it. Otherwise, you may have to chunk it. Either way is fine by me. I prefer that you use the whole file.
3. To save money, test your code on a small fraction of the file (say 5% of the tokens) and once it works reliably, then you may use the full files.
4. Use code in Homework 3 modules to adapt to this homework. You will develop a Python/LLM based solution for each data file.
5. Validate the retrieval queries against manual inspection to ensure that similarity search works well. Make observations on the validation effort.
6. Push the queries to fall outside the document domain and see what the answers are. Bulletproof your solution using one or more of the techniques we discussed in class to ward off such red-teaming attacks. Make observations on the resiliency of your solutions.
7. You have a lot of options here: changing the language models, changing embeddings, and how much red-teaming attacks you want to perform. But the more you try, the more expensive it might get but also the better your solutions will be. You may use less expensive or free models to reduce the cost. There is no need to use GPUs for this assignment.
8. Also, it would be of interest to upload the documents via the chatbots (ChatGPT or Claude or Gemini) to see if the answers are similar to what you are getting via API calls.
9. Also, whether the chatbots exhibit similar vulnerabilities to red teaming and the techniques you used in the API for bulletproofing work here as well.
10. I want you to do the best work possible here – lots of experimentation will help you learn things really well.
11. **Submission: Do not submit data files. I have them.**
12. Your code files.
13. Your narrative on red-teaming, what techniques you tried, what you observed and how successful you were in warding off attacks.
14. Your narrative on what models and embeddings worked and how much you spent would also be of interest to me for the entire assignment vis a vis the number of tokens you used.
15. I would love to see your work on items 8 and 9 and a narrative on it as part of your submission.