**PART 3 : Critical Thinking**

**Task 8 :**

* I used the "pdfplumber" library of python to extract text from pdf. pdfplumber is a powerful library that allows for easy extraction of text, data and images from PDFs, making it a valuable tool for data analysis and automation tasks we can use regular expression (RegExp ) to find particular text or string from extracted data.
* Many research papers contains diagram and images so according to me it is the best library we can use to extract text from pdf.
* I used “tiktoken” library of python for divide the text into tokens.

**Task 9 :**

For writing a prompt i follow this steps:

* First we specify the characters in the prompt so that the openai understands what type of response he has to give.
* After that describe about the previous process
* Than provide all the inputs required for task
* Ater that describe the openai’s task
* At the end define the rules and precautions so the openai provide us a accurate response
* For the task that doesn’t required too much reasoning, just describe the task.

**Task 10 :**

I faced many challenges during this task

* For task-2 i tried to use the “spaCy” library of NLP, but i found difficulties to get accurate response so i used openai to perform this task. Same i used openai for task-3 to search title and author from topic.To perform task-4(Convert the extracted fields into a structured JSON file) i used regular expression, but it did not provide the accurate results we need.
* When i used openai for task-4 than i faced the issue related to maximum tokens than i used “tiktoken” library of python and convert the extracted text into tokens. I divided the response into chunks than at the end combined all the chunks and get response into json file.
* I tried to perform task-3 using openai but i can’t get accurate response so i used Faiss vectorstore of langchain. Faiss is a library for efficient similarity search and clustering of dense vectors. It contains algorithms that search in sets of vectors of any size, up to ones that possibly do not fit in RAM.

**Task 11 :**

If i repeat this task than i will try to do this task using langchain, LlamaIndex and openai, NLP etc. Because langchain and llamaindex provide more functionalities.