Paper I:

The study presents new MuRAG (Multimodal Retrieval-Augmented Generation) for PDFs with images and texts sourcing and maintaining their connection. The proposed methodology involves this process: the textual items are extracted from the documents and each page of the document is converted into a single image. Thus, encoding these images and linking summarized descriptions to each page-image creates embeddings that allow unifying the storage of image and textual information in a knowledge base. In the retrieval phase of the proposed system, cosine similarity is used with top k relevant text and image chunks from this knowledge base and is fuse with multimodal LLM for a better and coherent response. The effectiveness and efficiency of responses are investigated using hit rate and mean reciprocal rank of the retrieved information and system generated responses’ correctness, relevance, and fidelity using mean correctness, mean relevancy, and the sample faithfulness scores.

**Strengths:**

* The proposed MuRAG efficiently retains the link between images and text about as well as is important for documents containing interconnected figures and textual content.
* By putting the summaries in tags and assigning them images, the model also optimizes computations and makes the process easier.
* The approach is assessed in terms of the recall and precision and other measures concerning the quality of the response.

**Weaknesses:**

* Each page is encoded as a single image; it can be disadvantageous in the sense that on some pages there may be more than one image, and the response quality might be impacted.
* The solution is based on high-resource embeddings and their storage, which can be a concern for scaling to great amounts of data or texts of great length.
* The use of four datasets decreases the generalization of the results attained in performance testing to other possible document structures and search queries.

Based on the findings, it is clearly seen that with respect to relationship between image and text, the proposed MuRAG is superior to typical MuRAGs. The experiments conducted on four QA data sets (Short-form, Long-form, MCQ and T/F) demonstrate statute performances with better hit rate and mean reciprocal ranks. Response generation scores of correctness, relevancy, and faithfulness are also amplified because of better conformity with input queries. MuRAG always works similarly to show the scores of multiple models and we have shown how the proposed MuRAG can be generalized for different LLMs like OpenAI and Gemini LLMs although the proper LLM choice depends on different query type. Thus, the proposed MuRAG is indeed efficient for disparate document types, which consist of both images and written text, thus improving multimodal retrieval and generation.

Paper II:

This research used a text generation model from the Llama-2 family developed by Meta, known as the Llama-2-7b model. Fitted for its high capability of apprehending syntactic patterns, Llama-2-7b has been retrained by supervised labeled data to achieve a harmony of response relevant helpfulness and response safety. It is developed on transformer architecture, which makes use of self-attention capabilities to capture context within 4,096 tokens setting. This sequential approach helps to provide the appropriate output for such application based on question and answer, thus is effective in the sensitive, empathetic nature of the chatbot application.

The chatbot implementation process utilizes Retrieval-Augmented Generation (RAG) in order to select the information type preferred by her from the available sexual harassment resources such as legal rights, helplines, and NGOs in India. The data comes from PDF files and read and processed with the help of the PdfPlumber package, where the text is ‘taken out’ from the file. The pulled content is divided into reasonable pieces of text, in which all words are embedded into a 68-dimensional word embedding space using the pre-trained BERT model fine-tuned for this task. These embeddings are saved in ChromaDB to make it easy to access when the app comes into contact with a user.

When a user types a query, the chatbot starts the isolation step to fetch relevant context sensitive chunks from the vector DB by using a LangChain retriever. Both the retrieved content and the chat history are delivered to the Llama-2 model to match the user’s intention and improve the relevance and emotional tone of the response.

Next, in the Llama-2 model, the user query is tokenized, and based on the obtained context, a response is generated, and later, the tokens are translated into natural language. The prompt template defines the response tone to be followed while forming responses and includes factors such as empathy, respect, and sensitivity into the text. The existence of a suicide warning system into the chatbot architecture has been trained to detect any sign of negative emotion or that necessitates user self-harm information, and guiding the user to seek professional help when needed. The interaction dynamics are based on a script, which adheres to the principle of ‘do no harm’ to create the environment which promotes the victim’s autonomy.

The response accuracy for Llama-2 chatbot was above 95% with a trend observed for highly empathetic and non-stigmatizing messaging. It was able to consistently recognize various forms of harassment and give the correct resources while not considering gender or location. The model’s performance was also assessed based on how well it is equipped to address uncertainty. Whenever it didn’t have the information, it informed the users politely to seek professional help or when it was not sure, offered clarification.’

Interestingly, the chatbot in some cases returned less related helplines; the problem was especially profound for male victims, as the amount of data from some regions of India was insufficient. But the intended helpline was typically available and modifications are expected to be made to improve on these areas.

This work demonstrates how the Llama-2 LLM can further enhance chatbots for delicate use cases such as support for survivors of harassment. Despite demonstrated accuracy and sensitivity over 95% the model provides safe atmosphere for response. Further developments, for example, carrying out the web scraping in real-time for newer contact information, moving to cloud-based service system, and adding support for multiple languages would make it much easier to use and would transform the potential that AI can provide in assisting harassment victims.

Paper III:

This research aims to evaluate the impact of three primary AI strategies: Retriever-Augmented Generation (RAG), fine-tuning, and prompt engineering all of which are applied using their most suitable and appropriate approaches. The design of RAG comprises integration of a retrieval subsystem with a pre-trained Large Language Model, implemented with the help of the Langchain infrastructure. This integration makes it possible for the chatbot to use data contextual, in real time. Data was partitioned and organized, after which became vectors and saved in a vector DB called Milvus, from where the latest data relevant to the user input was retrieved. When fine-tuning, LLMs (LLaMA2 and Falcon) were trained to perform based on the criteria of conversational data in the “openassistant-guanaco” data set. The previous approach enabled domain-specific accuracy by tuning the model for relevance in response quality using the Hugging Face transformers library. Finally, in an application of effective input prompts called as ‘‘Prompt Engineering, ‘’LangChain ‘s ‘‘PromptTemplate’’ was used to develop triggering inquiries and manage previous conversation history for creating top quality responses (with almost negligeable training expenses).

The results underline advantages of each method which cannot be achieved by using other approach. Fine-tuning proved to be the most effective providing the greatest accuracy (87.8%) and the highest BLEU score as high as 0.81, the model demonstrated a high level of versatility to the requests of a specific domain and fluency in providing contextually appropriate answers. Perplexity, with the overall score in this model 10.3, gives clear evidence of a very refined conversational context and designation the present model as being the most effective in terms of response generation. The same can be said about RAG which yielded 84.5% of correct references of text and 0.76 on the BLEU scale at the same time RAG has also the advantage of using updated information which is quite crucial in real time applications. Although it had a slightly more complex PPL score of 11.50, RAG’s ability to provide an immediate and up-to-date reply improves chatbot applicability all the more for rapidly changing fields. While the accuracy of the Prompt engineering model was the lowest with 83.2% and BLEU of 0.74, the approach allowed an efficient and easily scalable method that required minimal retraining to be implemented in any limited resource setting.

They have drawbacks nonetheless: Although, the fine-tuning approach accurate, it demands a lot of resources and cannot be used where information is changing frequently, unlike the external points that we are using in this study. As RAG grows complex and manages higher levels of perplexity it appears to face occasional struggles with queries that are ambiguous or otherwise unexpected and it seems highly contingent on the quality and accessibility of third-party information for its responses. However, prompt engineering is cheaper and more flexible than fine-tuning but the approach fails to provide accurate adaptation. Thus, while tutorial or learning systems may do well for simple user interactions it may not do well for very complex regions or even highly dense regions of the interaction matrix, although improvements in the specificity of the prompting can lead to better results.

Therefore, it can be concluded that RAG, fine-tuning and the overall work on the prompt engineering all provides different value additions to improve the chatbot experience. Tuning refines the final responses for much higher relevancy and coherency especially if the specificity of the response area’s is paramount. The main strength of RAG is the ability to make a real-time query to the relevant data, which can significantly positively affect the effectiveness of its use in cases where only fresh information is suitable. While the figures demonstrate that prompt engineering has lower efficiency, the relatively low cost and high adaptability make it a reasonable solution for relatively fast and unsophisticated implementations. The research that may be done in the future could be to integrate RAG with fine-tuning to better understand the text and return more relevant information or develop more prompt engineering methods for greater depth of the conversation. They could widen the possibility of using AI chatbots in different real-life situations that will improve the interaction and efficiency of the task accomplish.