Talk To Document using Graph Rag and Knowledge Graph

Pratik Dabade[1], Ajinkya Masal[2], Shivani Lokhande[3], Pratik Narole[4], Prof.Pravin Kamble[5]

1Student, Dept. of IT, Trinity College of Engineering and Research, Pune
2Student, Dept. of IT, Trinity College of Engineering and Research, Pune
3Student, Dept. of IT, Trinity College of Engineering and Research, Pune
4Student, Dept. of IT, Trinity College of Engineering and Research, Pune
5Assistant Professor, Dept. of IT, Trinity College of Engineering and Research, Pune

ABSTRACT

In the current age of advanced technologies and numerous data processes, obtaining useful contents from long, unorganized pieces of writing is still very hard. The standard ways of searching do not work well most of the times such as the relevance and precision of the answers. Therefore, we suggest the existence of an intelligent chatbot system, which extends upon the existing methods by further imbibing several new technologies like Large Language Models (LLM), Graph RAG and Knowledge Graphs. In this system, users can upload large files and ask questions to receive precise answers based on the context. With the enhanced solution, people utilize LLM's to comprehend sophisticated queries and Knowledge Graphs to interlink the data in a logical way. The system also applies community detection, which involves organizing similar information into perspectives to help in the rapid processing of queries. This method weaves complex data and technology together with users in mind for better results when searching for information. It is an easier, faster, and more effective way to obtain relevant information across a variety of fields such as academic studies, medicine, and information management systems.

Keyword : *LLM(Large Language Models), RAG(Retrieval Augmented Generation), Knowledge Graphs.*

1. INTRODUCTION

As we live in an age where there is too much information available, there arises a need to access information from large volumes of text in an effective and timely manner. Traditional forms of search are not very effective in dealing with intricate and extensive data, which results in unnecessary hard work and loss of time. To help address these issues, we suggest the creation of a smart chatbot system using advanced modern apps such as Artificial Intelligence (AI) systems, Large Language Models (LLMs), augmentation technique Graph Retrieval augmented generation (RAG) and knowledge graphs.

The user-friendly design of this chatbot system allows a reader to upload great chunks of text and ask questions in natural language. With LLMs applied for query comprehension and a Knowledge Graph applied for the information organization found on the documents, the system achieves precision and relevancy in the answers fetched. The user experience when searching in large amounts of data is set to change for the best by making this system 'smarter', faster and most importantly easier to use when looking for information.

Additionally, Graph Retrieval Augmented Generation improves how information is understood and the context in which it is related when retrieving the most suitable information needed. This creates a situation where the answer does not just zone in on certain areas of the document but there is a zone of focus that answers the question in a more thorough manner, resulting in the possibility of better and much further reaching answers. This is more advantageous in the context of large unstructured data sets where providing the relevant context is key in answering a question.

This research intends to prove the relevance of the system to delivery of services in various environments; from that of helping students do research to getting information in the healthcare sector. Due to its ability to analyze and mine out vital information from large datasets, the chatbot system can be of extreme importance in sectors which deal with lots of information and require quick access to specific pieces of information.

2. PROBLRM STATEMENT

To build an intelligent bot system in which users can speedily retrieve information from a large document by asking questions and the system responds with the relevant extracted content. With the integration of sophisticated AI technologies such as the use of Large Language Models (LLM's), the responses are relevant, prompt and within the context of the material, reducing complexity in information management and time utilization.

3. OBJECTIVES

Develop a chatbot interface that allows users to ask natural language questions and retrieve accurate, concise answers from large documents using LLMs. Build a Knowledge Graph and implement efficient indexing to detect communities and identify clusters of related information for faster query processing. Generate hierarchical community summaries that condense the key insights from each detected community in the knowledge graph. Leverage these summaries to efficiently answer user queries by retrieving contextually relevant information based on the graph structure.

4. LITERATURE SURVEY

significant insights and reveals the relative advantages of the designed applications. Zheng Tai et.al [1] Many tools for designing, developing, and administrating distance and blended courses in e-learning are available today. This paper presents the concept of an Integrated Development Environment (IDE) for e-learning, aiming to demonstrate its benefits and some of its main features.In the e-learning course development process, an author may use a number of different applications before finally bringing together weakly connected outcomes in one virtual space. As for educational purposes, usually some content management systems coupled with a number of authoring tools are implemented. Instructional design is a systematic approach which involves a rigorous analysis of learner characteristics, learning objectives, and context of learning (i.e. any constraints) before the actual design of instruction. The technological advances of today should not only focus on development of content and corresponding methods for students. This longitudinal study adopts an integrated approach by comparing curriculum development frameworks with their implementation in a blended e-learning environment. Moore's theory of transactional distance is examined in relation to curriculum development using e-learning.

Praveen Bansal et.al [2] The emergence of remote and blended mentoring techniques for educational purposes leads to consideration of digital platforms and social media for the actual practice with advanced pros and cons. The objective of the research work is to substantiate theoretical underpinnings of innovative processes organization management through the development of recommendations on the application of the structural-dynamic approach to creating comprehensive enterprise innovative processes development management systems.For instance, geographical exploration is considerably challenging without first selecting a dataset and thereby narrowing down the focus of the study. But this is not the case in KE. Structured knowledge –good quality, well organized and formally articulated knowledge- affects worker s effectiveness. Most self-diagnosis systems are prone to over-utilization or disregard of important clinical information.

Christian Weber et al. [3] In the growing field of personalized education, it has become more and more vital to provide justifications for the learning recommendations made to focus on their content. The recent emergence of Large Language Models (LLMs) and generative AI has made it possible to create human-like justifications for these recommended actions. Their precision is still lagging, however, notably due to the delicate nature of the field referred to as education. Therefore, we present a strategy that incorporates knowledge graphs (KGs) into the instructions of an LLM, to serve as the factual infinity for the instruction and minimize the chances of hallucinations as well as keeping appropriate details within the learning environment. Through the use of KGs, where certain knowledge relations are stored, learning recommendations which fit with the aims of the learner

are provided respectively. The explanations are based on text templates which are then filled in by LLMs and completed with the help of relevant experts. In our research, the learner's involvement ensured that the high level of information coherence and relevance was maintained all through the process of prompt design. We compared it the approach with the classical metrics evaluating lexical similarity – Rouge-N and Rouge-L as well as with qualitative evaluation from both experts and learners. The differences in favor of Recall and Precision were significant, allowing even in terms of filling out out the explanations – text obtained from pure GPT models – and containing much less information that could be disproved.

Zhongjian Hu et al.[4] have done some prior work using LLMs like GPT-3 in knowledge-based Visual Question Answering (VQA) tasks. However, we believe that knowledge incorporation can greatly improve the capabilities of LLMs for reasoning tasks. LLMs have been applied to a number of tasks, and more specifically, LLM performance has been enhanced through the use of knowledge graphs.

Yucheng Shi et.al [5] Chowdhury et al. (2022) described ChatGPT, A new state-of-the-art LLM achieved amazing results in various NLP tasks. Irrespective of its remarkable success, this system can only be useful in the real world to a limited extent due to two key reasons: it cannot be readily fine-tuned for the specific downstream tasks, and is very opaque in the way it reaches a conclusion. In doing so, we suggest a new framework which utilizes the strength of ChatGPT but focuses on its interpretable aspects for purposes such as text classification. Our framework undertakes a process of knowledge graph construction where ChatGPT is used to cleanse and organize raw data into a structured form. This information is then organized in a graph format, which is subsequently applied to fit an explainable linear classifier for the prediction task. This methodology is tested by conducting experiments on four benchmark dataset addresses and the performance is compared with the performance obtained in simply using ChatGPT models for text classification. The findings indicate that this approach is considerably more effective in terms of the performance of the predictions than using ChatGPT directly for the classification of the texts, and provides more clarity in how conclusions were reached compared to methods of the earlier times.

5. PROPOSED SYSTEM

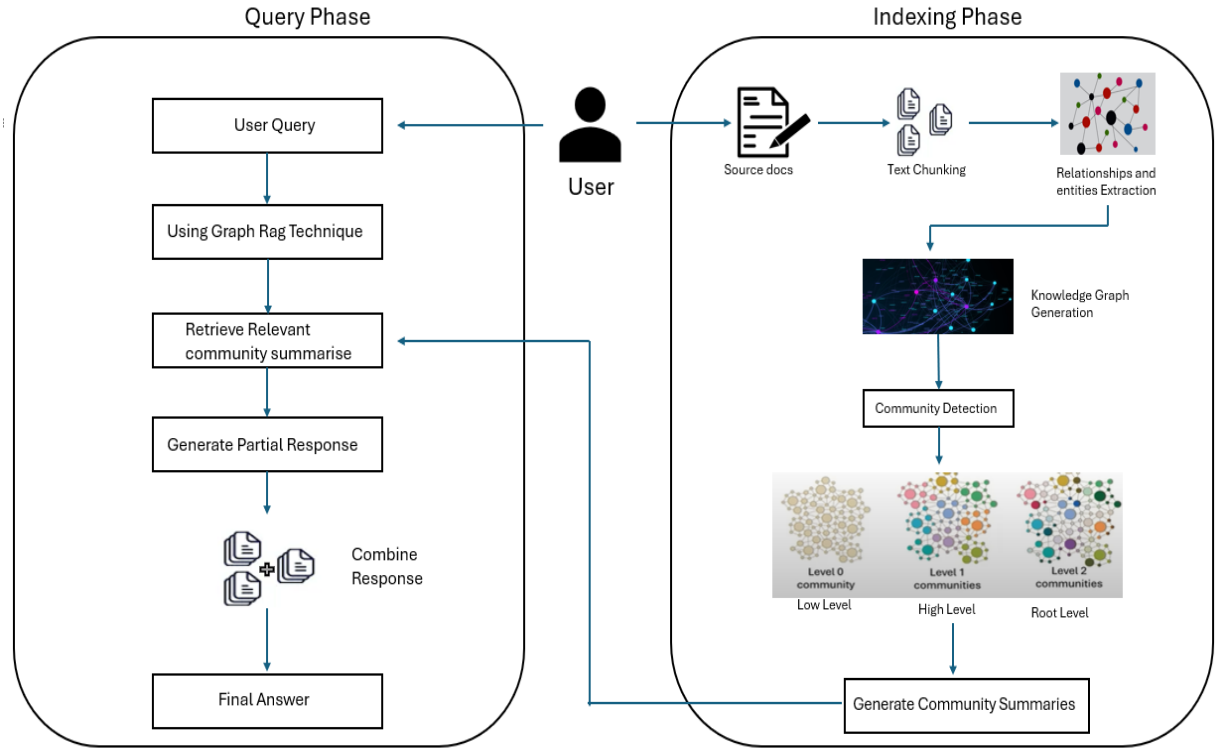


Figure 5.1: Architecture of proposed system

Everything starts with acquiring the source materials which can be considered the primary data of the system. These can include any type of information, be it text, images, or any other medium pertinent to the subject matter in question. The clarity and depth of these materials affect information retrieval systems, which makes them all the more important to include appropriate high-quality source materials.

After these source documents are acquired, the system applies text chunking, which is aimed at partitioning large volumes of text into smaller and more convenient pieces. This chunking up of the information enables the system to manage the information better; hence making it possible to index the information and fast retrieve the specific information when a user query is made. Use of text segmentation enables the system to seek relevant information within pieces of text rather than entire documents, thereby improving the efficiency and precision of the response.

In every process that follows text chunking takes place, the system identifies relations and entities from the text bundles. This entails recognizing important ideas, phrases, and their relationships within a given body of text. By this means of describing relations, a system building that database which will contribute to the conveinient and purposeful search of the information at latter stages is obtained.

The relationships and entities that have been extracted are subsequently structured in a Knowledge Graph. This graph is an organized view of the information that indicates how one entity relates to another. The Knowledge Graph does this by assisting in the improvement of information search and the quality of the answers produced by the system in response to user requests, as it clearly illustrates how various data points are interconnected.

After the Knowledge Graph is made, community detection algorithms help to cluster the information with the assistance of a graph. These algorithms check the graph and are able to locate clusters of data points with common properties or themes. Better yet, if such information is already organized in communities, the system can retrieve it more rapidly because relevant contents will all be placed in one area and made available during the query.

After the community detection, the system will produce summaries for the detected communities. The purpose of these summaries is because they bear the main points within each settlements making them easy to understand the contents without delving into details. In that sense, the system resorts into summarizing these communities because whenever there are queries, it needs to go into these communities and pick appropriate information only for the users appreciation of the content.

LLM'S can also be incorporated in the indexing phase to improve the relationships and entity extraction. Therein, the system is able to leverage on the power of the LLMs and analyze larger blocks of texts in a much better way, which helps to reveal relations and contexts that are missed in the considered approach. And this gives the Knowledge Graph richer content and higher precision, which in turn leads to better retrieval results.

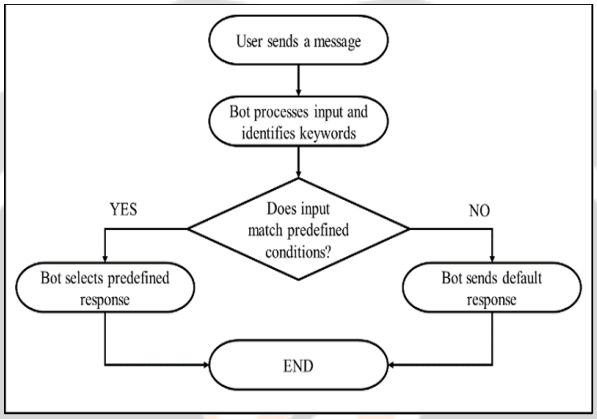
The process of Query Phase commences when the end-user enters a query, which can be in the form of a question or an information request. The system decodes this query to understand the user's purpose and the information that he or she is looking for. This first stage is critical to make certain that the following retrieval process is consistent with the user's requirements.

After the user query has been comprehended, the system employs the use of the Graph Retrieval Augmented Generation (RAG) technique. This technique improves the retrieval aspect by making use of relationships and structures present in the Knowledge Graph. With the enhancement provided by the RAG technique, the system is able to understand the context of the query better, and as a result, is able to find the data most appropriate to the user's request.

Once the query is prepared using the Graph RAG technique, the system produces summaries from the applicable communities that were found during the indexing. This phase guarantees that the information presented to the user is precise and relevant, utilizing the brief knowledge of communities with respect to the user's question.

According to the acquired community overviews, the system formulates an incomplete answer. Such an answer contains a few relevant information about the user's query which is not yet complete. The creation of intermediate response allows the system to pull some information from different community overviews and

present a conclusive answer later on. The snipped answers within different community overviews are blended together to form a single and complete one. This happens so that all the important information is incorporated and a satisfactory answer to the customer's question is provided. The process of Convergence Additivity enriches the final output by bringing together different strands of thought and information in the final answer. Eventually, the user receives the ultimate response. This answer is presented in a clear manner and correlates to the question the user has asked in a straightforward way. The aim is to make the last output not only corresponds to the expectations of the user but also provides the user with a nicely structured information on the issue at hand. During the Query Phase, LLMs are central in formulating the final response, in general and in that particular case. The and the last community summaries are processed using LLMs to obtain a reasonable, coherent and contextually proper answer to the question. Smarter language models, who can process and respond to subtleties in language, enhance the experience of the target user while responding to the user's requests. In addition, LLMs are query type and format tolerant which helps in handling various request from users easily.



Activity Diagram

The use case described here presents the various steps that take place in the chatbot while engaging a user, processing their input, and giving appropriate replies according to certain conditions. To begin with, a user encodes a message, which in turn activates the bot's message processing stage aimed at retrieving key details of the message, or keywords. This is the first step in the message processing chain that helps the bot in determining what the user is trying to accomplish as useful information is collected to help the bot know how to react to the user.

The next step after the extraction of keywords is when the system logic attempts to determine if the input can be appropriately categorized by any conditions preset by the bot. Should the conditions be applicable, the bot chooses and delivers a prepared answer that corresponds with the user's message. When the conditions do not apply, the bot's answer will be a standard one which could be a request for additional information or a simple response. This systemic approach guarantees that the bot is able to manage both anticipated responses and those that are out of its classification in a seamless manner thus enables the users to have a coherent experience. The phase is over after the bot's output has been rendered in the form of a response.

6. RESUITS

The system being presented uses knowledge graphs and the Graph Retrieval Augmented Generation (RAG) technique, making information retrieval more accurate and relevant. The burden of information is lifted from the user by community summaries that suit their needs and the content generated by Large Language Models (LLMs), which increases user engagement. This performance is evaluated using quantitative metrics of accuracy, precision, and recall statistics, over response time, and the system's user satisfaction surveys assess qualitatively the relevance of the generated answers. In addition, the generation of a knowledge graph on the fly will help the system incorporate new information and continue to be relevant and useful. The structure is designed to be scalable, permitting the surfeit of data input and various types of queries to be managed. Probing the integration of LLMs and

retrieval systems is noted as a future research direction promising 'big' benefits in education, healthcare, customer support, and so on. All in all, such system marks progress in information retrieval and proves to be beneficial for the fields that seek precise and timely information.

7. CONCLUSION

To sum up, the chatbot system that has been put forward significantly improves information retrieval with the use of Large Language Models still known as LLMs, Retrieval Augmented Generation known as RAG and in Knowledge Graphs. This enhances the capabilities of retrieving large amounts of contextually relevant information from very big texts both in terms of accuracy and speed. The complication in the user interaction has been reduced and short community summaries have been provided thus the system is more user friendly and efficient. The dynamic knowledge graph makes it amenable to new information hence it is useful in many areas.

8.FUTURE SCOPE

The anticipated growth of the intelligent chatbot system proposed, which combines large language models, graph retrieval augmented generation and knowledge graphs, is immense and disruptive. It is capable of changing the face of information retrieval in various sectors, including, but not limited to, health care, legal research, academia and business intelligence, by providing accurate, context-relevant responses to large complex datasets. The system in the future may be enhanced to accommodate multimodal data, user-centric approaches and real time knowledge but also addressing the language and field barriers. Accelerated progress on privacy, security and responsible AI will make it possible to use the solution in wider contexts with confidence. The users can expect the system to become more integrated within corporate systems as well as improve in its ability to manage information that is complex and constantly changing. This can help remove information access barriers, enhance use of technology to facilitate advanced learning and transform industries by enhancing access to decision support systems.

9. REFERENCES

1. Junfeng Zhang, Yang Zhang, Minnan Chu, Shun Yang, Taolei Zu: "A LLM-Based Simulation Scenario Aided Generation Method " 2023 IEEE (ITOEC) | 979-8-3503-3421-0.

2. Hasan Abu-Rasheed , Christian Weber ,Madjid Fathi : "Knowledge Graphs as Context Sources for LLM- Based Explanations of Learning Recommendations" 2024 IEEE(EDUCON).

3. Yucheng Shi ,Hehuan Ma, Wenliang Zhong,Qiaoyu Tan, Gengchen Mai, Xiang Li, Tianming Liu, Junzhou Huang: ChatGraph: Interpretable Text Classification by Converting ChatGPT Knowledge to Graphs" 2023 IEEE International Conference on Data Mining Workshops (ICDMW).

4. Yucheng Zhongjian Hu, Peng Yang*, Fengyuan Liu, Yuan Meng, and Xingyu Liu: "Prompting Large Language Models with Knowledge-Injection for Knowledge-Based Visual Question Answering" Big Data Mining and Analytics, September 2024.

5. Yucheng Shi (2023) "ChatGraph: Interpretable Text Classification by Converting ChatGPT Knowledge to Graphs."

6. Daeseung Park (2024)" AStudy on Performance Improvement of Prompt Engineering for Generative AI with a Large Language Model**."**

7. Alexander Tobias Neumann (2024An LLM-Driven Chatbot in Higher Education for Databases and Information Systems."

8. Ramaswami Mohandoss(2024)" Context-Based Semantic Caching for LLM Application ."