IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 10, April 2025
Impact Factor: 7.67

Transforming Website Navigation and Interpretation with a Retrieval Augmented Generation Chatbot

Dr. Lovely Mutneja, Aniket Dabhade, Yash Dudhe, Aachal Dandale, Sakshi Chopde
Prof Ram Meghe College of Engineering and Management, Badnera, India
dabhadeaniket97@gmail.com, yashdudhe414@gmail.com
achaldandale123@gmail.com, e-mail:chopadesakshi125@gmail.com

Abstract: *With the exponential growth of digital content, website navigation and information retrieval have become increasingly complex for users. Traditional search mechanisms often fail to provide precise and context-aware results, leading to inefficiencies in browsing. This paper presents a novel approach to website navigation and interpretation using a Retrieval-Augmented Generation (RAG) chatbot integrated with a Large Language Model (LLM). The chatbot leverages retrieval-based techniques to extract relevant website content while utilizing generative AI to enhance user interaction and query resolution. Implemented using Streamlit, the proposed system efficiently decodes website structures and provides intuitive responses, reducing user effort in information discovery. Our evaluation demonstrates that the chatbot significantly improves navigation efficiency and accuracy compared to conventional website search functionalities. This research contributes to the advancement of AI-driven website accessibility and presents opportunities for further enhancements in web-based information retrieval.*

Keywords: LLM, Llama-2, chatbot, Navigation , Interpretation

I. INTRODUCTION

The rapid expansion of the internet has led to an overwhelming volume of digital content, making website navigation and information retrieval increasingly challenging for users. Traditional search functionalities, such as keyword-based searches and menu-driven navigation, often fail to deliver precise, context-aware results, leading to inefficiencies and frustration. As websites grow in complexity, there is a growing need for intelligent systems that can enhance user experience by providing quick and accurate access to relevant information..

Recent advancements in artificial intelligence (AI) and natural language processing (NLP) have enabled the development of Retrieval-Augmented Generation (RAG) models, which combine retrieval-based search mechanisms with the generative capabilities of Large Language Models (LLMs). These models can effectively retrieve relevant content from structured and unstructured data sources while generating natural and coherent responses to user queries.

II. RELATED WORK

Website navigation and information retrieval have been extensively studied in the fields of information retrieval (IR), natural language processing (NLP), and artificial intelligence (AI). Traditional search engines and website indexing techniques have long been used to help users find relevant content. However, these approaches rely primarily on keyword-based matching, which often leads to irrelevant results due to the lack of contextual understanding.

A. Traditional Website Search and Navigation

Conventional website search mechanisms use full-text search engines and structured query techniques such as SQL-based filtering or Boolean search models [1]. While these methods efficiently retrieve structured data, they struggle with unstructured web content, making them less effective for websites with dynamic or large-scale information. Moreover, traditional menu-based and hierarchical navigation systems require users to manually explore multiple pages to find relevant information, increasing cognitive load [2].

DOI: 10.48175/IJARSCT-25692

B. AI-Driven Search and Conversational Interfaces

With the advancement of AI, researchers have explored chatbot-based solutions to enhance user experience in web navigation. Early chatbot systems primarily relied on rule-based or retrieval-based models, which offered predefined responses based on keyword matches or decision trees [3]. However, these systems lacked flexibility and adaptability to user-specific queries.

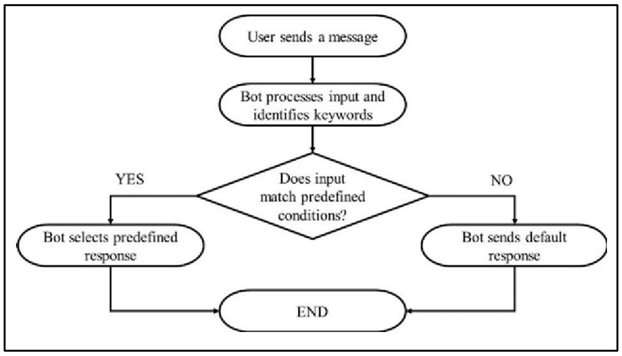

User sends a message
Bot processes input and identifies keywords
Does input match predefined conditions?
YES
NO
Bot selects predefined response
Bot sends default response
END

Fig.1 Workflow of a rule based chatbot

chatbot Recent breakthroughs in transformer-based models, such as BERT [4] and GPT [5], have improved search and conversational AI by enabling contextual understanding and dynamic response generation. Studies have demonstrated that LLM-powered chatbots outperform traditional search methods by providing more contextually relevant and human-like interactions [6].

C. Retrieval-Augmented Generation (RAG) for Information Retrieval

The Retrieval-Augmented Generation (RAG) framework, introduced by researchers at Facebook AI [7], combines retrieval-based and generation-based approaches to improve information accuracy and relevance. Unlike standalone generative models, RAG retrieves relevant documents before generating responses, ensuring that chatbot outputs are both factually grounded and context-aware. Prior work has shown that integrating RAG into chatbots enhances their ability to provide precise and reliable information, especially in domains requiring real-time data interpretation [8].

D. Streamlit for AI-Powered Web Applications

Streamlit has emerged as a popular framework for deploying interactive machine learning (ML) and AI applications [9]. Its lightweight architecture enables seamless integration of LLM-based chatbots into web interfaces. Research has demonstrated that Streamlit-based AI solutions can significantly improve user engagement and accessibility by offering real-time interactions and intuitive visualizations [10].

E. Research Gap and Contribution

Despite significant progress in AI-driven search and navigation, there is still a gap in efficient website interpretation using RAG and LLMs. Existing solutions either lack real-time retrieval capabilities or fail to generate contextually relevant responses. This paper addresses these limitations by proposing a Streamlit-based RAG chatbot that enables intelligent website navigation, improving information accessibility and user experience.

ole in supporting survivors of sexual harassment, offering a scalable and accessible platform for assistance and empowerment.

III. METHODOLOGY

A. Model architecture.

This study uses the Llama-2-7b model which is part of the Llama-2 family of pre-trained generative text models created by Meta and released in July 2023. They use an auto- regressive approach - which sequentially generates text based on the context provided to the model. We use the Llama-2-7b model since it is capable of handling complex linguistics since it has been pre-trained on a vast corpus of about 2 trillion tokens of text data and supports longer context lengths of up to 4096 tokens. For dialogue cases, the model has been fine- tuned in a supervised manner - using labelled data. The model is a great resource to use for the use case of this study since, it is built to balance both helpfulness and safety in its responses. In most cases, the Llama-2-7b model has been found to outperform most of its closed-source competitors such as ChatGPT [14].

The model is built on the powerful transformer Architecture model. The transformer model was introduced in 2017 and relies on self-attention techniques to capture contextual information from the input text. It consists of an encoder-decoder structure, where the encoder processes input tokens and the decoder generates output tokens. The multi- head attention mechanism allows the model to attend to different parts of the input sequence simultaneously, enhancing its ability to learn complex patterns [14].

B. Workflow

The chatbot created for this study works by using a systematic workflow to generate responses based on user queries. It uses Retrieval-Augmented-Generation (RAG) to retrieve relevant information from the store. The workflow begins with PDF files that contain information about resources and laws pertaining to sexual harassment. The text is extracted from these pdf files with the help of the PdfPlumber library [15].

The text extracted is then split into smaller chunks and converted into vector representation. This embedding of text into numerical vector form helps the Llama-2 model understand the underlying semantic meaning of the text. Which in turn, helps retrieve relevant contexts from the Vector database during answering. This embedding of chunks is done with the help of the fine-tuned BERT-base model, trained on the MS MARCO dataset [16]. It maps the text into a 68-dimensional dense vector space and is hosted on the HuggingFace community [17]. This process is done only once.

While using the model for question and answering scenarios, users submit a query to the model, triggering a search process for relevant contexts or information in the ChromaDB. These contexts are sent to the Llama-2 LLM, which processes these contexts and generates a meaningful and human-understandable response. This response will be formatted based on the contexts as well as user intent (which is obtained from the query). This high-level workflow can be seen in Fig. 2 below.

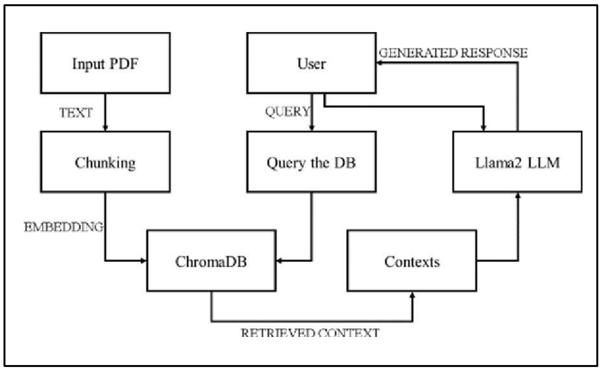


Fig. 2 Llama-2-7b RAG chatbot

C. Retrieval Augmented Generation

The Retrieval-Augmented Generation (RAG) framework enhances and further boosts the responding capability of the chatbot created. Traditionally, chatbots rely on pre-trained models or rules that generate responses that lack contextually appropriate information or user-specific responses. RAG battles this limitation by allowing the chatbot to dynamically retrieve relevant and up-to-date contexts from the knowledge base. By providing more detailed and reliable information, RAG significantly enhances user engagement and trust in chatbot interactions.

The model created uses a LangChain [18] retriever to retrieve the most relevant contexts from the vector database. It accepts a string query as input and returns a list of the top-n context embeddings from the ChromaDB vector database.

The selection of these documents is based on their similarity to the input query. Cosine similarity is used to check the similarity between the contexts and the query. It calculates the cosine of the angle between two vectors (shown in Eq. (1))

$$\cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} \qquad (1)$$

where A and B are the numerical vector representations of the two texts.

When a query is given as input, the retriever calculates the cosine similarity between the query's embedding and the embeddings of all the documents in the database. The documents with the highest cosine similarity are considered the most similar to the query and the top-n is returned as the output.

D. Response generation

The Llama-2 model generates responses by first tokenizing the input text query, generating the response, and finally decoding the tokenized response. Tokenizing is the process of transforming text into numerical vector representation. Once the query is tokenized, it is passed to the LLM's 'generate' method which uses the model's learned parameters to generate a sequence of tokens from that response. This is then finally converted back to natural language form. This process allows Llama-2 to generate responses that are contextually appropriate and semantically coherent.

IV. IMPLEMENTATION

A. Website Data Extraction

Web content is scraped using BeautifulSoup or Scrapy, extracting key text information from HTML pages.

The extracted content is preprocessed (removing unnecessary elements like scripts, ads, and duplicate content).

Stopword removal: Unnecessary common words are filtered.

Whitespace and punctuation cleanup: Ensures clean text representation.

Text normalization: Converts uppercase to lowercase, removes special characters.

Sentence segmentation: Splits paragraphs into meaningful text chunks.

Duplicate removal: Avoids redundant data in search retrieval.

B. Preprocessing and embedding

The study employs a preprocessing technique to enhance the quality of the extracted textual data that will be then converted to vector embeddings. The first step was to remove any special characters and symbols that may introduce noise or inconsistencies. Stopwords were removed, and the large document was broken down into smaller and manageable chunks. For all of these processes, we use Spacy's stopword list, the transformer's StoppingCriteria and, the LangChain RecursiveTextSplitter.

The chunks are thus ensured to be complete sentences that aren't split med-sentence and that the vector embeddings capture the semantic information well. Using the 'msmarco-bert-base-dot-v5' model, we generate embeddings for each chunk. the embeddings created are then stored in a vector store (ChromaDB).

C. Creating the Llama-2 model

The Llama-2-7b Large language model used in this study was downloaded and loaded on the system. The model was fed with a template that ensured that the chatbot would respond with safe, helpful, sensitive, and empathetic responses. The template plays a huge role as a structured guide for generating responses in a consistent and context-aware manner.

It emphasises the importance of context – since this is crucial in generating relevant and helpful answers for the user query. The tone of the template is specifically set for empathy, respect, support, and non-judgement. The model is explicitly instructed not to generate harmful, inappropriate, or false content. Recognising the limitation of a chatbot in the case of counselling and long-term help, the model is instructed to guide and encourage the victim to seek professional help. It also looks at chat history to ensure that the user's need is fully understood and that the responses generated are relevant to their query.

The prompt template used for the model of this study is given below:

"""Use the following pieces of context to answer the question at the end. {context}. You are a helpful, respectful, and empathetic friend. Your friend has undergone a traumatic event and you are trying to help them process the whole event and take the next steps. You must be understanding and never blame them. Always answer as helpfully as possible, while being safe. And try to gradually guide them to a therapist after ensuring that they are in a safe space. Your answers should not include any harmful, unethical, racist, sexist, toxic, dangerous, or illegal content. Please ensure that your responses are socially and gender unbiased. Always answer using the context provided and refrain from answering on your own. If a question does not make any sense, or is not factually coherent, explain why instead of answering something not correct and ask for further explanation. If the context provided does not have any relevant information or if you don't know the answer to a question, please don't share any false information, reply by saying "Sorry, I am not sure of the answer - but please don't hesitate to contact a counsellor". You can also look into chat history. {chat_history}
Question: {question} Answer:"""

The relevant contexts are retrieved from the vector database and are passed to the model as {context} as seen in the prompt along with the chat history and user query which are temporarily stored in the memory.

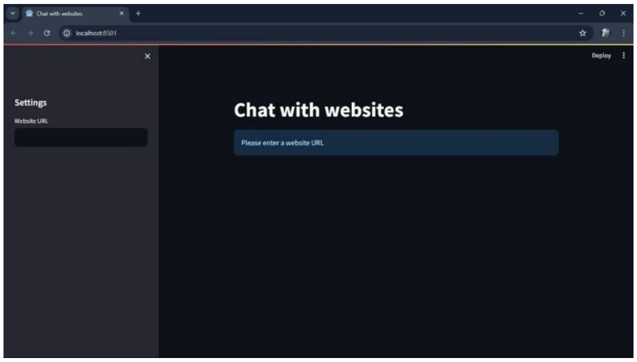
D. Response validation

The responses of the chatbot are consistently empathetic and non-judgemental. It acknowledges user emotions provides support and does not generate answers that accuse the victim of any wrong. In fact, the model always seems to be reassuring and kind. The model can handle cases of uncertainty quite well – and gently advises the user to seek professional help or provide more context.

The chatbot is fully competent at recognising the different types of harassment and tailoring its responses, accordingly, offering relevant advice or resources. It does not discriminate against the victim based on location, gender or identity and offers support regardless. All these results can be seen in Fig. 3 to 7.

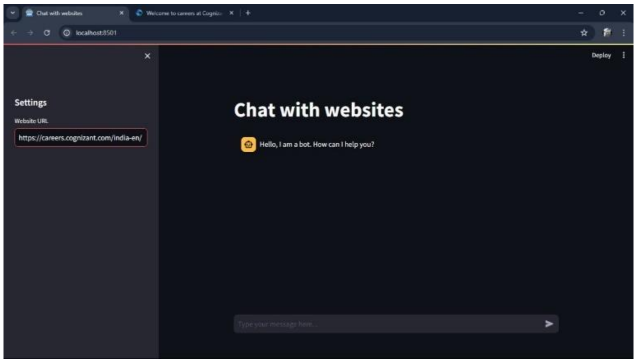
However, the chatbot is subject to sometimes retrieving the wrong helpline for 'boys' (Shown in Fig. 8). This could be due to the high difference in the amount of information available for men in India who have experienced sexual harassment and the very few helplines that are available for male victims in India. Nonetheless, it is worth noting that the right helpline also appears in the answer generated.

Moreover, the chatbot is completely hosted on the local machine, there is no issue of data privacy, but this can hinder the usage of the chatbot by multiple users. These issues must be considered when factoring in the further enhancements that could be applied to the model.



Please enter a website URL

Fig. 3 First Interface

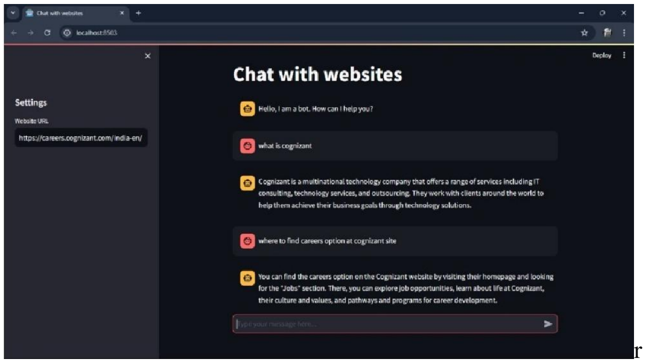


https://careers.cognizant.com/india-en/

Chat with websites

Hello, I am a bot. How can I help you?

Fig. 4 Site Processing



https://careers.cognizant.com/india-en/

Chat with websites

Hello, I am a bot. How can I help you?

what is cognizant

Cognizant is a multinational technology company that offers a range of services including IT consulting, technology services, and outsourcing. They work with clients around the world to help them achieve their business goals through technology solutions.

where to find careers option at cognizant site

You can find the careers option on the Cognizant website by visiting their homepage and looking for the "Jobs" section. There, you can explore job opportunities, learn about life at Cognizant, their culture and values, and pathways and programs for career development.

r

Fig. 5 Navigation and Interpretation output

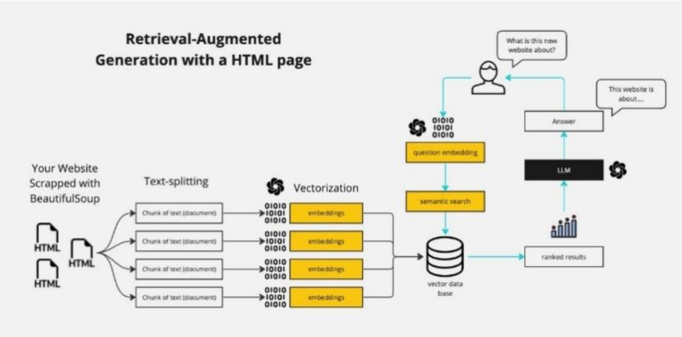


Fig. 6 Block Digram

V. CONCLUSION

In Conclusion, Our project successfully demonstrates the power of Retrieval-Augmented Generation (RAG) in enhancing website navigation and user interaction.By integrating retrieval-based search with LLM-powered responses, the chatbot provides contextually relevant and precise information, reducing search time and improving user experience.

This system bridges the gap between users and complex web structures, making information retrieval more intuitive and accessible.

Future enhancements may include multimodal support, voice-based interactions, and real-time website adaptation to further refine user experience.This approach represents a significant step forward in transforming how users interact with digital content, setting a new standard for intelligent website navigation., and accessible..

DOI: 10.48175/IJARSCT-25692

REFERENCES

[1] World Population Review. “Rape Statistics by Country.” https://worldpopulationreview.com/country-rankings/rape-statistics-by-country [Accessed: Feb. 18, 2024].

[2] Nasrin Sultana, “India Inc Sees Alarmingly High Unresolved Sexual Harassment Cases At Workplace,” Forbes India, Oct. 17, 2023 https://www.forbesindia.com/article/take-one-big-story- of-the-day/india-inc-sees-alarmingly-high-unresolved- sexual-harassment-cases-at-workplace/89043/1 [Accessed: Feb. 18, 2024].

[3] Karami, Amir, et al. "A systematic literature review of sexual harassment studies with text mining." Sustainability 13.12 (2021): 6589.

[4] Shechory-Bitton, Mally, and Liza Zvi. "Is it harassment? Perceptions of sexual harassment among lawyers and undergraduate students." Frontiers in psychology 11 (2020): 1793.

[5] Følstad, Asbjørn, et al. "Future directions for chatbot research: an interdisciplinary research agenda." Computing 103.12 (2021): 2915-2942.

[6] Yao, Yifan, et al. "A survey on large language model (llm) security and privacy: The good, the bad, and the ugly." arXiv preprint arXiv:2312.02003 (2023).

[7] Sorin, Vera, et al. "Large language models (llms) and empathy-a systematic review." medRxiv (2023): 2023-08

[8] Bauer, Tobias, et al. "# MeTooMaastricht: Building a chatbot to assist survivors of sexual harassment." Machine Learning and Knowledge Discovery in Databases: International Workshops of ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part I. Springer International Publishing, 2020.

[9] Li, Cheng, et al. "Large language models understand and can be enhanced by emotional stimuli." arXiv preprint arXiv:2307.11760 (2023).

[10] Rathnayaka, Prabod, et al. "A mental health chatbot with cognitive skills for personalised behavioural activation and remote health monitoring." Sensors 22.10 (2022): 3653.

[11] “rAInbow: Chatbot to Support Victims of Domestic Abuse.” World Justice Challenge 2021. rAInbow: Chatbot to Support Victims of Domestic Abuse | World Justice Project [Accessed: Feb. 18, 2024].

[12] Socatiyanurak, Vorada, et al. "Law-u: Legal guidance through artificial intelligence chatbot for sexual violence victims and survivors." IEEE Access 9 (2021): 131440-131461.

[13] Maeng, Wookjae, and Joonhwan Lee. "Designing a chatbot for survivors of sexual violence: Exploratory study for hybrid approach combining rule-based chatbot and ml-based chatbot." Asian CHI Symposium 2021. 2021.

[14] “Llama 2: Open Foundation and Fine-Tuned Chat Models,” HuggingFace, meta-llama/Llama-2-7b • Hugging Face [Accessed: Feb. 18, 2024].

[15] "pdfplumber." PyPI, version 0.10.4, released on Feb 10, 2024. https://pypi.org/project/pdfplumber/ . [Accessed: Feb. 18, 2024].

[16] Hugging Face. "sentence-transformers/msmarco-bert- base-dot-v5." Hugging Face, 2024. https://huggingface.co/sentence-transformers/msmarco- bert-base-dot-v5 . [Accessed: Feb. 18, 2024].

[17] Hugging Face. "The AI community building the future." Hugging Face, 2024. https://huggingface.co/ . [Accessed: Feb. 18, 2024].

[18] "LangChain." LangChain, 2024. https://www.langchain.com/ . [Accessed: Feb. 18, 2024].

[19] O. Banerji, “Domestic violence helpline numbers, counselling and how to report cases smoothly,” IPleaders, Mar. 23, 2022. https://blog.ipleaders.in/domestic-violence-helpline- numbers-counselling-and-how-to-report-cases/ . [Accessed: Feb. 18, 2024].

[20] R. Baruah, “The Law Against Sexual Harassment,” Legal Services India, https://www.legalservicesindia.com/article/2545/The- Law-Against-Sexual-Harassment.html#:~:text=The%20Sexual%20Harassme nt%20Against%20Women,as%20well%20as%20international%20level [Accessed: Feb. 18, 2024]

[21] National Commission for Women, “Helplines,” National Commission for Women, http://www.ncw.nic.in/helplines [Accessed: Feb. 18, 2024]

[22] S. Goenka, “Sexual Harassment”, Legal Service India, https://www.legalserviceindia.com/legal/article-1323-sexual-harassment.html [Accessed: Feb. 18, 2024]

[23] “Home | Men Welfare Trust”. Men Welfare Trust, http://www.menwelfare.in/ [Accessed: Feb. 18, 2024]

[24] Saumya Uniyal and Siddhesh Surve. “Top 11
Organizations in India that Help in Cases of Molestation, Sexual Abuse & Violence”. TimesNext https://timesnext.com/top-organizations-in-india-that-help-in-cases-of-molestation-sexual-abuse-violence/#:~:text=Top%2011%20Organizations%20in%20India%20for%20Reporting%20Cases, %208%208.%20Azad%20Foundation%20 %20More%20items[Accessed: Feb. 18, 2024]

[25] Jahnvimehta. “Sexual Violence against Men in India”. Legal Service India, https://www.legalserviceindia.com /legal/article-4685- sexual-violence-against-men-in-india.html [Accessed: Feb. 18, 2024]

DOI: 10.48175/IJARSCT-25692