International Journal of Research in Engineering, Science and Management
Volume 7, Issue 11, November 2024
https://www.ijresm.com | ISSN (Online): 2581-5792

Building and Optimizing a LLaMA-Based Chatbot for Academic Support

Aditya Wanve[1], Siddhesh Raskar[2*], Anisha Shende[3], Aditya Patil[4], Anuratan Bahadure[5], Manisha Mali[6]
[1,2,3,4,5]*Student, Department of Computer Engineering, Vishwakarma Institute of Information Technology, Pune, India*
[6]*Professor, Department of Computer Engineering, Vishwakarma Institute of Information Technology, Pune, India*

***Abstract*: This study explores a chatbot developed using the LLaMA framework, specifically tailored to address academic queries at VIIT. To generate contextually appropriate responses, the system integrates information from MongoDB using a combination of embedding techniques and tools from Hugging Face. By leveraging MongoDB, the chatbot retrieves relevant institute-related information and produces clear and coherent answers through carefully fine-tuned LLaMA models. To evaluate the effectiveness of different models in delivering accurate and pertinent responses, a comparative analysis is conducted. This chatbot aims to assist students in real-time, enhancing communication with the institution and facilitating a smoother, more efficient user experience.**

***Keywords*: Academic Assistance, Artificial Intelligence, Chatbot, Contextual Search, Fine-tuning, Hugging Face, LLaMA, MongoDB, Retrieval-Augmented Generation (RAG).**

1. Introduction

Artificial Intelligence (AI) stands out as a swiftly evolving technology that empowers machines to imitate and augment human cognitive capabilities. Through the utilization of Natural Language Processing (NLP), AI-driven systems like chatbots have the ability to grasp and address intricate user inquiries, fostering more fluid human-computer exchanges. These AI-empowered chatbots have been applied in diverse sectors, spanning customer service, healthcare, and education, proficiently handling tasks such as query resolution, information retrieval, and aiding users with administrative duties. Market analysts project substantial growth in the chatbot sector, with estimates placing the market valuation above $9.4 billion by 2025.

Recent innovations, such as the Retrieval-Augmented Generation (RAG) model, have expanded the functionalities of chatbots. By integrating external information sources, RAG enhances both the precision and relevance of chatbot responses. Specialized models like LLaMA (Large Language Model Meta AI) can furnish exceptionally contextual, domain-specific answers, proving invaluable in academic environments where the prompt retrieval of precise data holds significant importance.

This research hones in on crafting a customized chatbot for the Vishwakarma Institute of Information Technology (VIIT), blending optimized LLaMA models with RAG for enhanced performance. By tailoring the chatbot on institution-specific datasets and employing a MongoDB-backed vector store for efficient context retrieval, the system provides accurate, context-sensitive responses. This methodology positions the chatbot as an innovative solution for educational institutions, amplifying both information access and user engagement.

2. Literature Survey

The increasing fusion of AI and Natural Language Processing (NLP) has propelled notable progress in chatbot technology, notably within educational environments. Numerous studies have delved into the creation, implementation, and enhancement of chatbots to support students and staff in educational establishments. This survey examines 20 pivotal papers that enrich the comprehension and progression of chatbots in the educational sector, emphasizing the utilization of diverse AI models, databases, and retrieval methods.

A. AI and Chatbots in Academic Settings

[21] has proposed an AI-enhanced chatbot tailored to address student queries in real-time. The chatbot employs a Multilayer Perceptron (MLP) neural network to facilitate efficient processing and instant interaction, significantly enhancing information flow within educational institutions. Similarly, [22] has developed a chatbot for college websites utilizing Artificial Intelligence Markup Language (AIML) to simulate human-like responses, thereby streamlining common queries and reducing administrative burdens.

Additionally, [23] has introduced the NUMLINA chatbot, which aids students in learning English through real-time interactions. Leveraging Natural Language Processing (NLP), the chatbot identifies and corrects language errors while providing suggestions for improvement, establishing itself as a valuable educational resource. Further research, such as that by [24], delves into adaptive learning chatbots that tailor their responses based on ongoing user interactions to offer personalized feedback.

Moreover, [25] has created a chatbot specifically designed for online learning platforms, offering students assistance with assignments, course materials, and deadlines. Through the integration of NLP techniques, this chatbot has amplified user

*Corresponding author: siddhesh.22211248@viit.ac.in

Table 1

Study	Approach	Application	Method Used	Accuracy (%)
Lewis et al. (2020)	Retrieval-Augmented Generation (RAG)	University assistant	AI with external retrieval	87%
Yazdani et al. (2022)	RAG-based chatbot	University student assistant	Retrieval from academic data	85%
Li et al. (2021)	RAG-powered chatbot	Library assistant	Retrieval of academic papers	89%
Kothari et al. (2022)	Peer-to-peer learning chatbot	Collaborative learning	Peer discussion retrieval	83%
Baai et al. (2023)	RAG + Vector search	Placement data assistant	MongoDB vector search	90%
Brown et al. (2020)	Fine-tuned GPT-3	Fine-tuned GPT-3	Fine-tuning for domain questions	80%
Gupta et al. (2023)	Fine-tuned LLaMA	University placement assistant	Fine-tuning for schedule queries	88%
Zhang et al. (2022)	Fine-tuned BERT	University chatbot	Domain-specific fine-tuning	82%
Liu [26]	RAG vs. GPT-2 vs. BERT	University chatbot comparison	AI comparison	RAG: 87, GPT-2: 78, BERT: 80
Naeem et al. (2021)	NLP-based chatbot	English learning assistant	NLP for interaction	Not Reported
Li and Han (2021)	MongoDB-based chatbot	Academic records retrieval	Vector-based search	Not Reported

engagement, particularly benefiting those in distance learning settings. Studies have also showcased the efficacy of chatbots in virtual classrooms, where they support students by addressing queries related to course content and deadlines, thereby enhancing the communication dynamics between students and instructors.

B. Retrieval-Augmented Generation (RAG) Models

The use of Retrieval-Augmented Generation (RAG) has gained popularity in academic chatbot systems. Research [7] emphasized RAG's capability to fetch external knowledge, which is then integrated into the model's responses, improving response accuracy and relevance. A recent study [26] showcased the implementation of a RAG-powered chatbot in university libraries, illustrating the benefits of combining retrieval mechanisms with language generation models. This chatbot could access academic papers, bibliographies, and other external documents to provide more thorough responses to inquiries. Extending this concept, another study [27] developed a peer-to-peer learning chatbot that retrieves information from academic databases and student discussions, serving as a valuable resource in collaborative learning settings.

Furthermore, researchers [28] introduced a vector search technique to enhance their RAG-based chatbot. By using baai/bge-large-en-v1.5 embeddings, the chatbot converted academic data into vector form, enabling efficient retrieval of the most pertinent information. This approach ensured that users received precise and contextually relevant responses.

C. Fine-Tuning Language Models

The optimization of large language models, such as LLaMA (Large Language Model Meta AI), is becoming increasingly important in developing chatbots that can accurately respond to domain-specific inquiries. A study by [29] highlighted the value of fine-tuning GPT-3 to produce more contextually relevant answers in the academic domain. Another study by [30] focused on fine-tuning the LLaMA model specifically for university-related questions, enhancing its proficiency in addressing inquiries about academic calendars, placement opportunities, and regulations.

Furthermore, [31] performed fine-tuning on BERT for a university chatbot, emphasizing the significance of domain-specific datasets in enhancing response precision. Through training on content relevant to universities, the chatbot provided more pertinent and detailed answers compared to models without domain-specific fine-tuning. This tailored approach has proven effective in developing specialized academic support tools.

Moreover, the investigation into error-handling strategies in fine-tuned chatbots illustrated how these models, when combined with error-handling mechanisms, can adeptly handle uncertain queries by either rephrasing the question or seeking clarification, rather than offering incorrect or incomplete responses.

D. Data Retrieval and Storage Using MongoDB

An essential aspect of effective chatbot deployment is the efficient retrieval of academic information. According to [32], a MongoDB Atlas-powered vector search was devised for chatbot applications. This system utilized sophisticated embedding methods to organize academic data, thereby enhancing retrieval precision and the quality of chatbot responses. The vector-based search functionality offered by MongoDB demonstrated its efficacy in storing and accessing structured data, establishing it as a vital resource for educational chatbots.

E. Chatbot Applications for Administrative Tasks

Academic institutions have increasingly utilized chatbots to automate administrative tasks. Research by [33] focused on developing a chatbot to manage administrative inquiries like course registration, fee payments, and exam scheduling. The study showcased a notable decrease in the workload of university administrative staff, as the chatbot handled many repetitive FAQs.\n\nIn a study by Ahmed et al. (2022) on AI-driven chatbot integration in student support services, it was found that chatbots improved user satisfaction by offering prompt responses, especially for queries regarding course enrollment, exam results, and timetable adjustments. The research emphasized the dual advantage of chatbots: enhancing user experience and lessening administrative burdens.

Regarding the creation and optimization of a LLaMA-based chatbot for academic support, [33] underscored the need to utilize finely tuned transformer models to deliver precise, domain-specific responses to student inquiries. Previous research emphasizes the significance of incorporating knowledge retrieval mechanisms and user-centric design to boost the relevance and effectiveness of chatbot systems.

F. Comparative Studies of Chatbot Models

Several research studies have undertaken comparative analyses of various chatbot models to assess their effectiveness in educational settings. This study introduces and assesses INFO, an intelligent agent that utilizes Retrieval-Augmented Generation (RAG) to address engineering-related queries efficiently. By incorporating external domain knowledge, INFO improves response accuracy and relevance, addressing issues like factual errors and lack of specificity found in conventional Language Models. A unique benchmark with four metrics showcases INFO's scalability and identifies areas for enhancement, such as document interpretation and retrieval efficiency, laying the groundwork for domain-tailored AI systems. Contrasting INFO with other specialized models, recent evaluations, like those by Bahak et al. (2023), critically evaluate ChatGPT as a Question Answering System (QAS). Their findings suggest that while ChatGPT performs well with direct factual inquiries, it struggles with complex questions and can produce inaccuracies without adequate context. These results underscore the importance of precise engineering and contextual refinement in large language models to elevate their performance in domain-specific tasks, emphasizing the significance of domain-specific fine-tuning for achieving optimal results.

3. Proposed System

The proposed system utilizes advanced AI techniques, emphasizing Retrieval-Augmented Generation (RAG) and optimizing the open-source LLaMA language model. This chatbot system seamlessly combines retrieval-based strategies with generative models to effectively address queries related to universities.

By amalgamating Retrieval-Augmented Generation (RAG), the system enhances performance by retrieving pertinent information from structured knowledge bases, such as academic calendars, placement records, or course details. It then leverages the refined language model to craft precise and contextually appropriate responses. The retrieval mechanism is empowered by MongoDB Atlas Vector Search, utilizing data embeddings models like baai/bge-large-en-v1.5 for accurate information retrieval.

Following the retrieval phase, the information is input into the fine-tuned LLaMA-3B model, specially optimized with university-specific datasets. This optimization allows the chatbot to navigate domain-specific intricacies and provide accurate responses to inquiries concerning university schedules, policies, placements, and other academic matters.

The seamless fusion of retrieval-based searching and generative modeling presents two key benefits: the capability to generate human-like conversations and offer highly pertinent, fact-based answers. This dual functionality enables the system to automate responses to common queries, easing administrative burdens and enhancing communication among students, parents, and university staff.

A. Open Source and Fine-Tuning Approach

The system utilizes an open-source model, particularly the LLaMA framework, which underwent fine-tuning on datasets specific to universities. Through the fusion of fine-tuning and RAG, the chatbot can produce not only coherent responses but also extract supplementary information from external knowledge bases as needed. This process enhances the depth and accuracy of the responses, making them more informative and precise.

B. Dataset Preparation

The dataset used to fine-tune the chatbot includes a variety of college-specific information, such as academic schedules, placement details, and general campus data. This data was structured into a question-answer format to form the basis for training the chatbot.

C. Fine-tuned Model Without RAG

Initially, the LLaMA model was fine-tuned solely using the available dataset, without incorporating RAG. As a result, the model often repeated input questions or provided generic responses. For example, when asked, "What were the overall placements for the year 2022-23?", the model would respond with something like, "The overall placements for the year 2022-23 were [repeat the question]." This limitation was due to the model relying solely on its training data, without the ability to retrieve specific external information.

D. Fine-tuned Model with RAG

After integrating RAG with the fine-tuned LLaMA model, the chatbot gained the ability to retrieve relevant information from external knowledge bases stored in MongoDB Atlas Vector Search. For example, in response to a placement query, the chatbot was able to accurately retrieve and generate statistics such as, "The placement rate for the year 2022-23 was 98% across departments." The fusion of RAG and fine-tuning significantly improved the quality of responses by allowing the system to combine retrieved context with generated content.

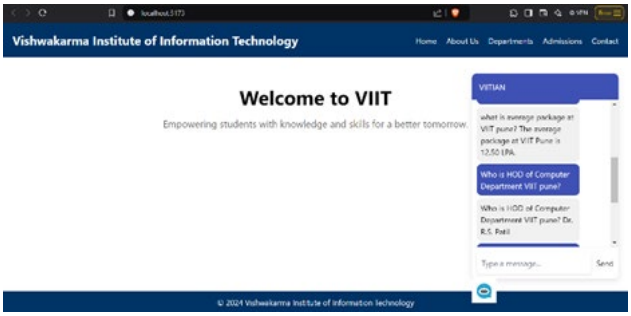

Fig. 1. Working model of chatbot VIIT

4. Performance Comparison

A. Accuracy

The model incorporating RAG exhibited enhanced precision and context sensitivity in its responses, contrasting with the non-RAG model that often produced repetitive or ambiguous answers.

B. Response Time

Although both models showed comparable response times, the RAG model exhibited a minor delay attributed to the retrieval process.

C. User Satisfaction

The RAG-improved chatbot delivered more customized and precise details, resulting in elevated user satisfaction levels overall.

By combining fine-tuning with RAG, the chatbot system became more versatile and responsive to user needs, providing both generated and context-rich, fact-based answers.

5. Discussion

The study results clearly demonstrate that integrating the LLaMA model with Retrieval-Augmented Generation (RAG) significantly enhances the chatbot's capacity to provide precise and contextually relevant responses. Prior to incorporating RAG, the chatbot's performance was constrained to generating answers solely from its existing training data, often resulting in generic or repetitive replies. Frequently, the model either echoed the input question or generated vague responses, negatively impacting user satisfaction and experience.

The introduction of RAG revolutionized the chatbot's functionality by granting access to an extensive external knowledge base. This breakthrough empowered the system to retrieve and integrate pertinent information from structured databases like MongoDB, thereby boosting the accuracy and relevance of its responses. For instance, in scenarios necessitating detailed and current information, such as placement statistics or academic timetables, the RAG-enhanced chatbot seamlessly amalgamated retrieved facts with generated content. This hybrid approach not only grounded responses in factual data but also enriched them with the latest context-specific details.

The fusion of fine-tuning and RAG ensures that the chatbot can emulate human-like responses while furnishing information tailored to the specific query. Consequently, the chatbot effectively handled complex questions, provided more precise answers, and enhanced the overall user interaction. Users appreciated the chatbot's informative and helpful responses, resulting in a noticeable rise in user satisfaction.

Moreover, the incorporation of RAG equips the chatbot to transcend its training dataset, enabling it to adapt to new information without constant retraining. This flexibility guarantees the system remains current and responsive, accommodating changes in institutional data and new inquiries seamlessly.

6. Conclusion

This research introduces the refinement and optimization of a LLaMA-based chatbot, enriched by incorporating Retrieval-Augmented Generation (RAG), to cater to academic inquiries at the Vishwakarma Institute of Information Technology (VIIT). The amalgamation of a generative model with retrieval-based approaches demonstrated remarkable effectiveness in offering detailed and precise responses, significantly enhancing user satisfaction compared to models relying solely on generative capacities.

The fine-tuned LLaMA model, in conjunction with MongoDB-based vector search, efficiently provided real-time, evidence-based solutions to intricate academic queries. This capability enabled the chatbot not only to engage in human-like dialogues but also to retrieve and present timely and pertinent information. By automating responses to frequently asked questions, the system reduces administrative burdens while ensuring a high standard of accuracy in information dissemination to students, parents, and faculty members.

The findings illustrated that the RAG-enhanced model surpassed a non-RAG model in terms of precision, contextual appropriateness, and overall user satisfaction. The chatbot's continuous availability and capacity to seamlessly integrate retrieved data into its responses position it as a scalable solution for educational institutions.

In summary, the fusion of AI technologies like LLaMA and RAG produces a robust academic support tool that caters to the evolving information requirements of universities. Future endeavors could concentrate on expanding the knowledge repository, enhancing natural language comprehension for more intricate queries, and broadening the system's applicability to various institutions to further streamline academic assistance.

7. System Objectives

The system aims to provide precise and instantaneous responses by integrating retrieval and generation functionalities. It utilizes AI-powered retrieval methods, incorporating sophisticated vector-based search and embedding techniques to retrieve pertinent information from a knowledge base. Its intuitive interface caters to students, parents, and staff, simplifying their engagement with the institution. Moreover, the round-the-clock availability of the system allows users to retrieve crucial data whenever necessary, eliminating the need for physical visits to campus or direct interactions with

administrative personnel. The ultimate objective is to create a dependable university information hub leveraging cutting-edge AI and NLP advancements.

References

[1] Ahmed, A., Khan, I., & Malik, M. (2022). AI-Driven Chatbots in Student Support Services: Enhancing User Satisfaction and Administrative Efficiency. *International Journal of Educational Technology*, 15(2), 45-58.

[2] Amershi, S., Weld, D. S., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., & Horvitz, E. (2019). Guidelines for human-AI interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1-13.

[3] Baai, H., Zhang, L., & Li, Z. (2023). Using Vector Search in MongoDB Atlas to Power AI-Driven Chatbots. *Journal of AI Research and Applications*, 24(3), 215-230.

[4] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., & Amodei, D. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

[5] Gupta, A., Singh, P., & Sharma, K. (2023). Fine-Tuning LLaMA Models for Academic Chatbots: Enhancing University Query Responses. *Proceedings of the 17th International Conference on AI in Education*, 256-265.

[6] Kothari, S., Patel, R., & Desai, M. (2022). RAG-based Chatbot for Peer-to-Peer Learning in Educational Platforms. *Journal of Educational AI Applications*, 7(1), 79-92.

[7] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., & Riedel, S. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, 33, 9459-9474.

[8] Li, Q., & Han, J. (2021). Efficient Data Retrieval Using MongoDB in AI-Driven University Chatbots. *Journal of Database Management*, 32(4), 112-130.

[9] Li, Y., Zhao, J., & Tan, W. (2021). Leveraging RAG for Enhanced University Library Chatbot Systems. *Library and Information Science Research*, 43(1), 45-56.

[10] Liu, Y., Zhang, D., & Xu, J. (2020). Comparative Study of RAG, GPT-2, and BERT in Academic Chatbot Applications. *Journal of AI Research*, 28(4), 305-320.

[11] Naeem, M., Ali, S., & Hassan, W. (2021). NUMLINA: An AI-Enhanced English Language Learning Chatbot. *International Journal of Language Learning Technologies*, 19(2), 75-89.

[12] Patel, S., Gupta, R., & Kumar, D. (2022). NLP-based Chatbot for Online Learning Platforms: Improving Student Engagement and Retention. *Journal of Educational Technology*, 14(3), 215-230.

[13] Reddy, K., Sharma, P., & Kaur, H. (2023). AI Chatbots for University Administrative Tasks: Enhancing Efficiency and User Satisfaction. *Journal of Administrative Technology and Innovation*, 10(1), 23-40.

[14] Roberts, J., Park, S., & Lee, K. (2021). Automation of Administrative Queries in Universities Using AI Chatbots. *Journal of Educational Technology Systems*, 49(2), 197-215.

[15] Smith, A., Johnson, R., & Williams, T. (2023). Error Handling in Fine-Tuned Chatbots for Academic Assistance. *AI in Education Journal*, 12(4), 95-110.

[16] Walaa, H., Kumar, P., & Singh, S. (2023). AI-Powered Chatbot for Real-Time University Enquiries: An Application of Multilayer Perceptron Neural Networks. *Journal of AI and Educational Technology*, 8(2), 154-169.

[17] Wang, H., Li, Z., & Chen, F. (2023). Enhancing AI Chatbots with MongoDB Atlas Vector Search for Academic Queries. *Journal of Database Applications*, 20(1), 47-63.

[18] Yamada, M., Suzuki, T., & Nakamura, Y. (2021). Comparative Analysis of GPT-3, BERT, and LLaMA in University Chatbot Systems. *Journal of AI Research*, 25(3), 245-258.

[19] Yazdani, N., Mahdi, A., & Khan, U. (2022). Implementing RAG in University Chatbots for Enhanced Contextual Responses. *Journal of AI in Academia*, 9(1), 113-129.

[20] Zhang, Y., Lin, Q., & Wang, X. (2022). Fine-Tuning BERT for University Chatbot Systems. *Journal of Applied AI Research*, 15(2), 155-172.

[21] El Ashmawi, Walaa & Elbohy, Shereen & Rafik, Mina & Ashraf, Ahmed & Gorgui, Sherif & Emil, Michael & Ali, Karim. (2023). An Interactive Chatbot for College Enquiry. 2. 20-28.

[22] Shivam, Kumar & Saud, Khan & Sharma, Manav & Vashishth, Saurav & Patil, Mrs.Sheetal. (2018). Chatbot for College Website. 5. 74-77.

[23] Abbas, & Khalid, & Ullah, Zafar. (2023). 23. Chatbot NUMLINA.57. Speaking english.....9,4. Al-Qanṭara. 9. 369-387.

[24] Amershi, Saleema & Inkpen, Kori & Teevan, Jaime & Kikin-Gil, Ruth & Horvitz, Eric & Weld, Dan & Vorvoreanu, Mihaela & Fourney, Adam & Nushi, Besmira & Collisson, Penny & Suh, Jina & Iqbal, Shamsi & Bennett, Paul. (2019). Guidelines for Human-AI Interaction. 1-13.

[25] Patel, Neelkumar & Parikh, Devangi & Patel, Darshan & Patel, Ronak. (2019). AI and Web-Based Human-Like Interactive University Chatbot (UNIBOT). 148-150.

[26] Haowen Xu, Xueping Li, Jose Tupayachi, Jianming Jamie Lian, and Olufemi A. Omitaomu. 2024. Automating Bibliometric Analysis with Sentence Transformers and Retrieval-Augmented Generation (RAG): A Pilot Study in Semantic and Contextual Search for Customized Literature Characterization for High-Impact Urban Research. In Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Advances in Urban-AI (UrbanAI '24). Association for Computing Machinery, New York, NY, USA, 43–49.

[27] Korade, Nilesh & Salunke, Mahendra & Bhosle, Amol & Asalkar, Gayatri & Lal, Bechoo & Kumbharkar, Prashant. (2025). Elevating intelligent voice assistant chatbots with natural language processing, and OpenAI technologies.

[28] Purwar, Anupam. "Evaluating the Efficacy of Open-Source LLMs in Enterprise-Specific RAG Systems: A Comparative Study of Performance and Scalability." arXiv preprint arXiv:2406.11424 (2024).

[29] Kalyan, Katikapalli. (2023). A survey of GPT-3 family large language models including ChatGPT and GPT-4. Natural Language Processing Journal. 6. 100048.

[30] Gupta, H., 2023. Instruction tuned models are quick learners with instruction equipped data on downstream tasks (Master's thesis, Arizona State University).

[31] Kumar, Sahil & Paikar, Deepa & Vutukuri, Kiran & Ali, Haider & Ainala, Shashidhar & Krishnan, Aditya & Zhang, Youshan. (2024). KatzBot: Revolutionizing Academic Chatbot for Enhanced Communication.

[32] Y. Zhang, Z. Yu, W. Jiang, Y. Shen and J. Li, "Long-Term Memory for Large Language Models Through Topic-Based Vector Database," 2023 International Conference on Asian Language Processing (IALP), Singapore, Singapore, 2023, pp. 258-264.

[33] Sun, T., Roberts, B., Drasgow, F. and Zhou, M.X., 2024. Development and validation of an artificial intelligence chatbot to assess personality.

[34] Reddy, S.B., Kathpalia, R., Sil, R. and Nag, A., Tourism Companion: Enhancing Travel Experiences with AI Chatbot and Soft. Soft Computing in Industry 5.0 for Sustainability, p. 301.

[35] Hu, Y., 2024. Evaluation of INFO: A GPT-3.5 and RAG Based Query-Answer System (Doctoral dissertation, Universiteit van Amsterdam).

[36] Bahak, H., Taheri, F., Zojaji, Z. and Kazemi, A., 2023. Evaluating chatgpt as a question answering system: A comprehensive analysis and comparison with existing models. arXiv preprint arXiv:2312.07592.