RAG Chatbot for Policy Document Enquiry and Feedback

Karthick Nag K S[1], Dr. T Vijaya Kumar[2]

[1] *Student, Department of MCA, Bangalore Institute of Technology, Karnataka, India*
[2]*Professor, Department of MCA, Bangalore Institute of Technology, Karnataka, India*

---***---

Abstract

In today's era of rapidly evolving digital ecosystems, accessing and understanding large, complex policy documents remains a major challenge for students, institutions, and stakeholders. Traditional manual search methods are inefficient, time-consuming, and prone to misinterpretation. To overcome this limitation, this paper presents a Retrieval-Augmented Generation (RAG) powered chatbot that integrates large language models (LLMs) with similarity-based retrieval mechanisms for precise and context-aware responses. By leveraging a vector store and embeddings for document chunking, the chatbot retrieves the most relevant sections of the policy and generates human-like answers using the Grok LLM. This hybrid approach ensures factual consistency, improves user accessibility, and provides real-time feedback collection. Experimental evaluation demonstrates that the RAG chatbot significantly improves query accuracy, reduces response time, and enhances user satisfaction. The system shows strong potential for scalable, cost-effective, and interactive policy assistance in academic and administrative domains.

Keywords— RAG, LLM, FAISS, Policy Document, Chatbot, Information Retrieval.

I. INTRODUCTION

Universities and organizations often rely on lengthy policy documents to govern academic regulations, examinations, grading systems, progression rules, and administrative procedures. These documents, while comprehensive, are typically hundreds of pages long and written in formal legal or academic language, which makes them difficult to interpret quickly. Students, faculty, and administrators frequently face challenges when they need immediate answers to specific queries, such as credit requirements, grading criteria, or internship rules. Traditional approaches—such as manually searching PDFs or browsing FAQs—are not only time-consuming but also fail to capture the semantic intent behind user queries.

In recent years, conversational AI has emerged as a promising solution to bridge the gap between users and complex institutional knowledge. However, traditional chatbots, whether rule-based or retrieval-only, fall short in providing contextually accurate and user-friendly answers. Rule-based bots are rigid and limited to pre-defined queries, while retrieval-only systems often present raw excerpts without coherent explanations. On the other hand, large language models (LLMs) like Groq and GPT have the ability to generate natural and context-rich answers.

To address these challenges, **Retrieval-Augmented Generation (RAG)** has emerged as a hybrid solution. By combining semantic retrieval with generative reasoning, RAG ensures that responses are both factually correct and linguistically fluent. In this framework, relevant sections of policy documents are retrieved from a vector database (FAISS) using embeddings, and then passed to an LLM (Grok) that generates accurate, human-like responses.

Our proposed **RAG Chatbot for Policy Document Enquiry and Feedback** builds upon this architecture. Unlike traditional systems, it not only provides instant, context-aware responses but also includes a feedback mechanism that allows users to rate or refine responses. This ensures a continuous cycle of improvement, enhancing both accuracy and user satisfaction. Moreover, the system is scalable to multiple domains (academic, administrative, or corporate) and adaptable to different document types.

By making policy documents conversationally accessible, this project has the potential to reduce confusion, improve institutional transparency, and save valuable time for students, staff, and decision-makers.

II. LITERATURE SURVEY

The rapid progress in artificial intelligence and natural language processing has reshaped how users access and interact with digital information. While policy documents and academic resources are now widely available online, their length and complexity make traditional keyword search inadequate. To address this, researchers have increasingly focused on **Retrieval-Augmented Generation (RAG)**, which combines information retrieval with large language models (LLMs) to deliver accurate and context-aware responses.

Kulkarni et al. [1] demonstrated how reinforcement learning can optimize RAG pipelines, reducing token costs while maintaining accuracy. Similarly, Akkiraju et al. [2] proposed the FACTS framework, highlighting freshness, architecture, cost, testing, and security as critical factors for scaling enterprise RAG chatbots.

In education, Antico et al. [3] introduced the Unimib Assistant, a RAG chatbot for university students, showing promise but noting retrieval limitations. Neupane et al. [4] developed BARKPLUG V.2, which achieved high factual accuracy (0.96) in campus support. Beyond academia, Freitas and Lotufo [5] applied RAG to retail via Retail-GPT, which improved shopping assistance but revealed risks of hallucinations and security issues.

Several works have focused on accessibility and personalization. Oreški and Vlahek [6] introduced Flowise AI to simplify chatbot deployment for universities, while Wang et al. [7] proposed UniMS-RAG, integrating retrieval and generation for higher personalization. Nguyen and Quan [8] presented URAG, a hybrid system combining rule-based responses with RAG for university admissions, balancing accuracy with cost efficiency. Khana et al. [9] evaluated Llama-2 and Mistral models in RAG setups, confirming their value for student services. In healthcare, Nayinzira and Adda [10] developed SentimentCareBot, integrating sentiment analysis with RAG to improve mental health chatbot interactions.

Overall, prior research confirms that **RAG enhances chatbot accuracy and usability across domains**, but persistent challenges—hallucinations, retrieval precision, and scalability—remain. These insights directly shaped the design of the proposed RAG-based chatbot for policy document enquiry and feedback.

III. EXISTING SYSTEM

Most universities today provide access to their academic regulations and policy documents through online portals in the form of downloadable PDFs, keyword-based search tools, or static FAQ sections. While these systems make information technically available, they often fail to deliver relevant, context-aware, and user-friendly responses. Students and faculty members must manually navigate lengthy documents or repeatedly refine search queries, which can lead to frustration, wasted time, and misinterpretation of crucial rules.

Traditional chatbots have been introduced in some cases, but they typically fall into two categories: **rule-based bots** and **retrieval-only systems**. Rule-based bots rely on pre-defined patterns and responses, making them rigid and incapable of handling unexpected queries. Retrieval-only systems can fetch text fragments from documents, but they often present answers without sufficient explanation or contextual clarity. Moreover, both approaches lack adaptability to user intent and cannot provide the interactive, conversational experience expected in modern AI solutions.

Another limitation of current systems is the **absence of feedback mechanisms**. Users may encounter unclear or unsatisfactory answers, yet there is no structured way to capture this input for system improvement. This lack of continuous learning prevents existing solutions from evolving over time.

Disadvantages of Existing Systems

- Manual search is time-consuming, inefficient, and prone to human error.
- Rule-based chatbots lack flexibility and cannot adapt to varied phrasing of user queries.
- Retrieval-only systems often provide fragmented or incomplete answers without context.
- No built-in mechanism to collect user feedback for improving the quality of responses.
- Limited personalization, as current systems cannot adapt answers based on user type (student, faculty, or administrator).

IV. PROPOSED SYSTEM

To address the limitations of current university policy access methods, the proposed solution introduces a Retrieval-Augmented Generation (RAG)-based Chatbot specifically designed for policy document enquiry and structured feedback collection. The system leverages the combined strengths of information retrieval and large language models to provide accurate, context-aware, and user-friendly answers. Unlike traditional approaches, this chatbot not only retrieves relevant content but also interprets and explains it in natural language, making policies more accessible to students, faculty, and administrators.

The architecture is organized into **three key layers**:

1. **Document Processing:**
   Policy documents are ingested and pre-processed through parsing and chunking. Each chunk is embedded using an open-source embedding model, and these vector representations are stored in **FAISS**, a high-performance similarity search index. This allows for fast and scalable retrieval across large collections of regulations.
2. **Query Handling:**
   When a user submits a query, it is transformed into an embedding and compared against the FAISS vector store. The most relevant chunks are retrieved based on semantic similarity, ensuring that the chatbot understands both the context and intent of the question.
3. **Response Generation:**
   The retrieved chunks are passed into the **Groq Large Language Model (LLM)**, which synthesizes the information into a coherent, conversational, and policy-compliant response. The model ensures that answers are both **factually grounded in the source documents** and **presented in an easy-to-understand manner**.

In addition, the system incorporates a **feedback collection mechanism**. Users can rate or comment on responses, and this feedback is stored in a dedicated database. Administrators can later review this data to assess chatbot performance, identify knowledge gaps, and continuously fine-tune the system for improved accuracy and reliability.

Advantages of the Proposed System

- Provides **accurate and context-aware** responses by grounding answers in official policy documents.
- Eliminates the need for **manual navigation and document search**, saving time and effort.
- Highly **scalable**, capable of supporting multiple policies across institutions.
- Incorporates a **feedback-driven learning loop**, ensuring continuous system enhancement.
- Improves **user engagement and trust** by delivering natural, conversational explanations.
- Offers potential for **future integration** with student portals, mobile applications, or cross-institutional policy databases.

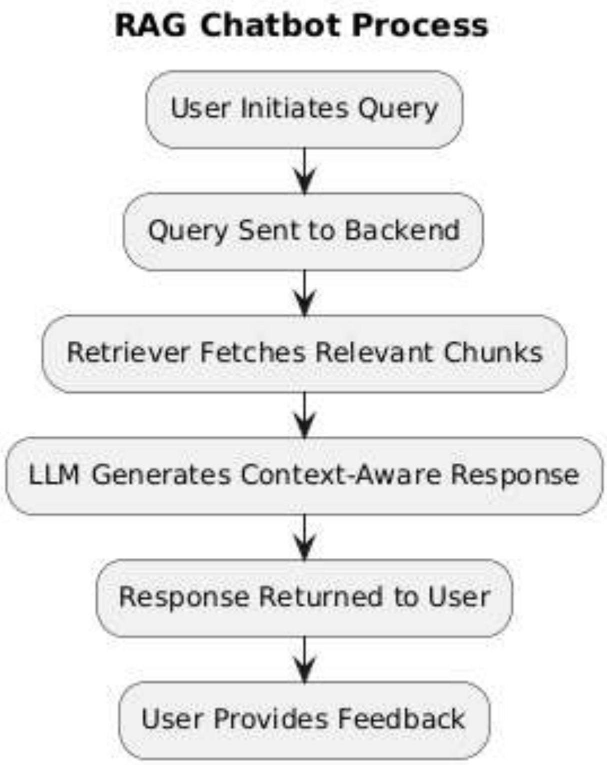
RAG Chatbot Process
User Initiates Query
Query Sent to Backend
Retriever Fetches Relevant Chunks
LLM Generates Context-Aware Response
Response Returned to User
User Provides Feedback

Fig 1: Proposed Model

V. IMPLEMENTATIONS

The implementation of the RAG-based Chatbot system is structured into four major components: **data preparation, backend, frontend, and database integration.** Each of these modules plays a critical role in ensuring smooth functionality, scalability, and accuracy of the chatbot.

- **Data Preparation:**
  Policy documents in PDF format are ingested using *PyPDF2*. The content is split into smaller, semantically meaningful chunks using *RecursiveCharacterTextSplitter*. Each chunk is then transformed into high-dimensional vector representations through an open-source embedding model. These embeddings are indexed and stored in *FAISS*, allowing for efficient similarity-based retrieval during query handling.

- **Backend:**
  The backend is implemented in *Python (Flask framework)*. It manages query processing, retrieval, and communication with the *Grok LLM API*. When a user submits a query, the backend uses FAISS to retrieve the most relevant document chunks and passes them to the Grok model, which generates a context-aware, natural language response. The backend also handles logging of queries and feedback, making the system adaptable and continuously improvable.

- **Frontend:**
  A *React-based web application* serves as the user interface. It provides an interactive platform where users can type queries, receive chatbot responses in real time, and submit ratings or feedback. The frontend ensures a smooth user experience with a responsive design suitable for desktops and mobile devices.

- **Database:**
  A *Microsoft SQL Server* database is used to store system metadata, user queries, and feedback records. This enables admins to analyse user interactions, track system performance, and enhance the model iteratively based on feedback.

Workflow:

1. The user enters a query in the frontend interface.
2. The backend retrieves relevant document chunks from FAISS.
3. The Grok LLM generates a natural, context-aware answer.
4. The response is displayed to the user in real time.
5. The user has the option to rate or provide feedback, which is logged in the database.

This modular implementation ensures that the chatbot is **scalable, efficient, and user-centric**, while also enabling future upgrades such as advanced analytics, additional policy integration, and improved personalization.

VI. DEPLOYMENT DIAGRAM

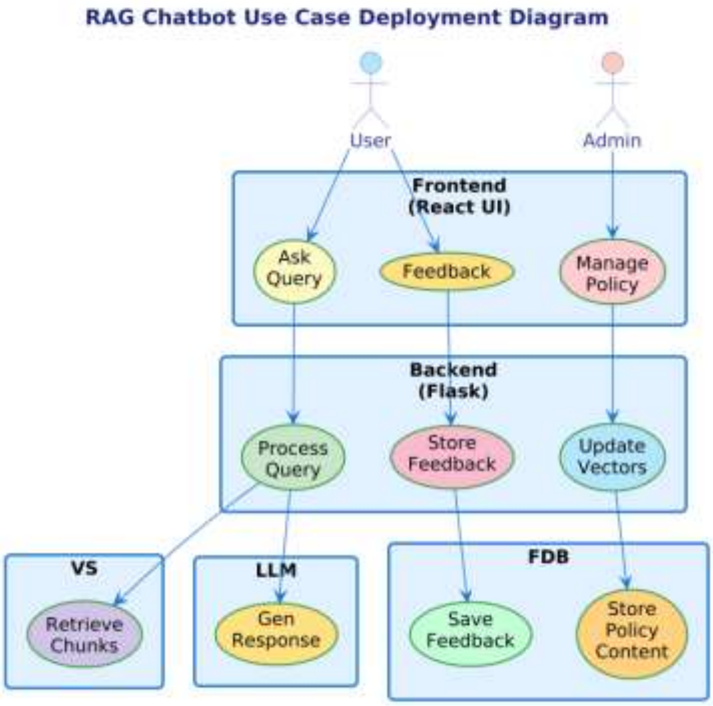

Fig 2: Use Case Deployment Diagram

VII. FLOW-CHART

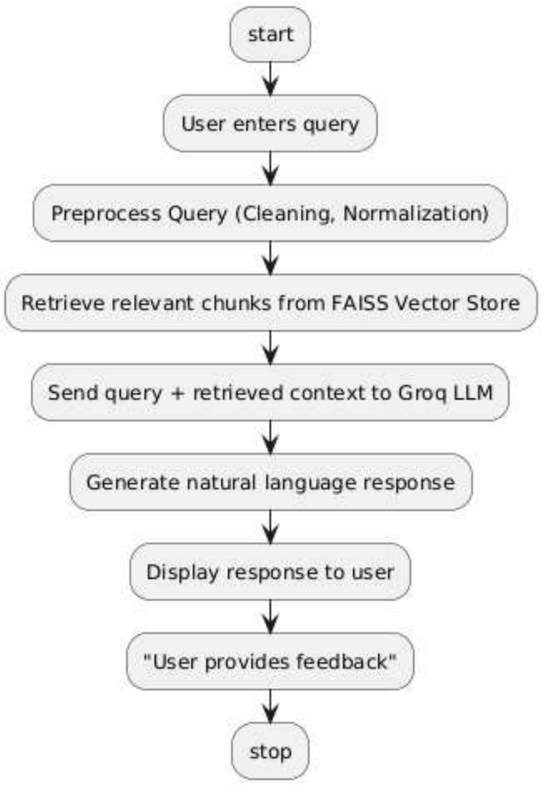
start
User enters query
Preprocess Query (Cleaning, Normalization)
Retrieve relevant chunks from FAISS Vector Store
Send query + retrieved context to Groq LLM
Generate natural language response
Display response to user
"User provides feedback"
stop

Fig 3: Flow-Chart

VIII. RESULTS AND DISCUSSION

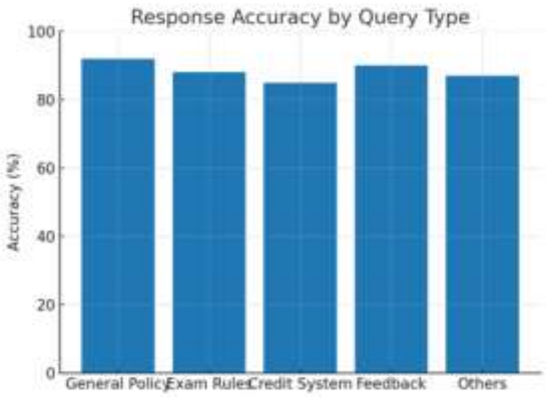

Fig 4: Response Accuracy Graph

This bar chart illustrates the response accuracy of the RAG chatbot by query type, measured as a percentage. The chart compares five query categories: General Policy, Exam Rules, Credit System, Feedback, and Others. The accuracy for each type is high, with General Policy queries achieving the highest accuracy (around 92%), while Exam Rules, Feedback, and Others each maintain accuracies above 85%. Credit System queries have the lowest, but still robust, accuracy in the mid-80% range. Overall, the graph demonstrates that the chatbot reliably delivers accurate responses across diverse policy-related queries, with only minor variations between different categories.

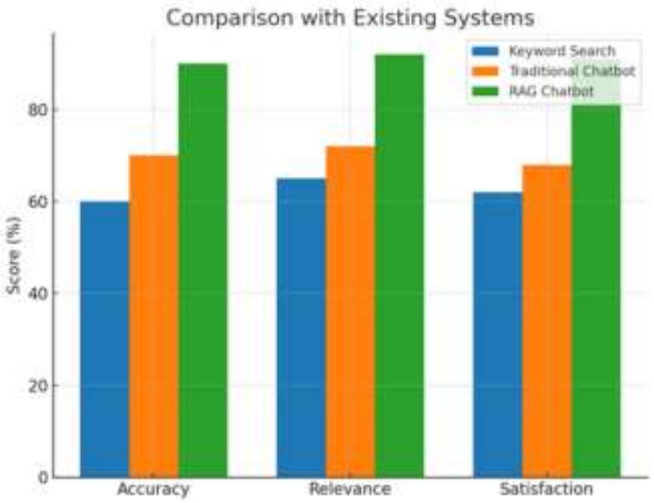

Fig 5: Comparison Graph

This bar chart compares the performance of Keyword Search, Traditional Chatbot, and RAG Chatbot systems across three metrics: Accuracy, Relevance, and Satisfaction. The RAG Chatbot consistently scores the highest in all categories, followed by the Traditional Chatbot, with Keyword Search showing the lowest scores. This indicates that the RAG Chatbot outperforms the other systems in delivering accurate, relevant answers and achieving higher user satisfaction.

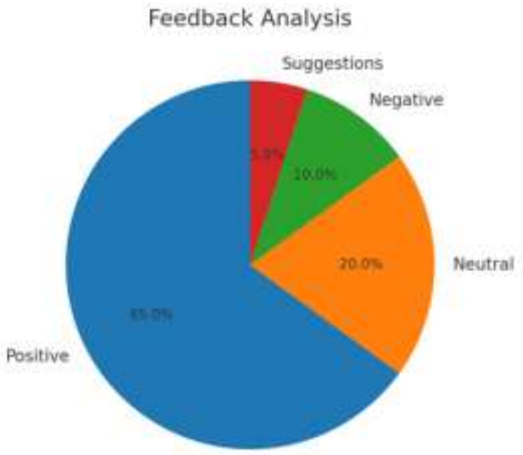

Fig 6: Feedback Analysis Pie-chart

This pie chart presents a feedback analysis for the RAG chatbot, breaking down user responses into four categories. The majority of feedback is positive, making up 65% of all responses. Neutral feedback accounts for 20%, negative feedback comprises 10%, and suggestions constitute the remaining 5%. This distribution indicates that most users have a favorable impression of the chatbot, with a smaller proportion offering neutral comments, criticism, or suggestions for improvement.

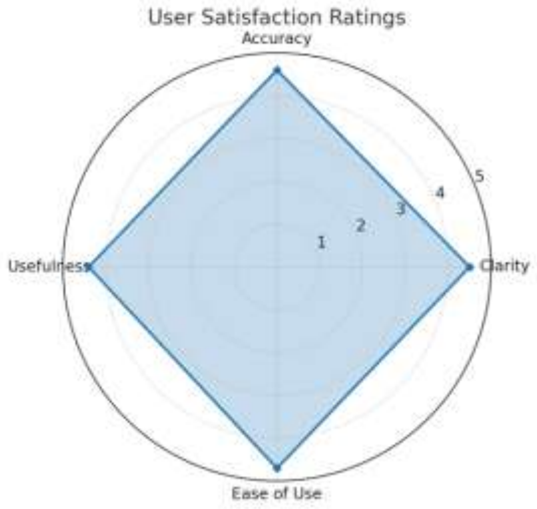

Fig 7: User Satisfaction Ratings

This radar chart displays user satisfaction ratings across four key aspects of the RAG chatbot: Accuracy, Clarity, Ease of Use, and Usefulness. All categories score highly, with each rating close to 4.5 out of 5. This indicates that users find the chatbot accurate, clear, easy to use, and useful, reflecting a strong overall user experience.

IX. CONCLUSIONS

This project successfully demonstrates a RAG-based chatbot designed to simplify access to complex university policy documents. By integrating **FAISS-based semantic retrieval** with **Groq's generative language model**, the system is capable of producing accurate, context-aware, and user-friendly responses. The use of document chunking and embeddings ensures that information retrieval is efficient and precise, even when users frame queries in diverse natural language forms.

An important feature of the system is the **feedback mechanism**, which allows continuous improvement through user evaluations of responses. This ensures adaptability over time and enables administrators to monitor performance, identify limitations, and refine the system further. Unlike traditional keyword search portals or static FAQs, the chatbot reduces manual effort, minimizes misunderstandings, and provides a more conversational and intuitive interface.

The system also demonstrates **scalability**—new policy documents can be added to the vector store with minimal effort, and the architecture can be extended to support multiple departments or institutions. From an institutional perspective, this contributes to improved transparency, faster resolution of student and staff queries, and overall enhancement of communication efficiency.

In summary, the RAG-based chatbot not only reduces query resolution time but also improves **information accessibility, institutional trust, and user satisfaction**. With planned extensions such as multilingual support, advanced analytics, and integration with mobile platforms, the system has the potential to become a **comprehensive and sustainable solution** for policy enquiry and feedback management in academic and administrative domains.

X. FUTURE ENHANCEMENTS

While the current system demonstrates strong performance in policy document retrieval and feedback management, several enhancements can be considered to further improve scalability, usability, and adaptability:

- **Mobile Application**: Extend support for Android and iOS platforms, enabling students and staff to access policies and submit queries on the go. Push notifications can keep users updated on policy changes, new circulars, or feedback responses.

- **Multilingual Support**: Incorporate machine translation and multilingual embeddings to support students from diverse linguistic backgrounds, ensuring inclusivity and accessibility.

- **Advanced Analytics**: Provide administrators with dashboards that track frequent queries, unresolved issues, and trends in feedback. This can help identify knowledge gaps, improve documentation, and optimize institutional communication strategies.

- **Integration with ERP/SIS**: Connect the chatbot with existing Enterprise Resource Planning (ERP) and Student Information Systems (SIS) to provide personalized responses based on user profiles (e.g., course details, semester, eligibility).

- **AI-based Feedback Analysis**: Use Natural Language Processing (NLP) and sentiment analysis to automatically categorize feedback as positive, negative, or neutral. This will allow institutions to take proactive actions to improve user satisfaction.

- **Voice-based Interaction**: Enable speech-to-text and text-to-speech features for users who prefer conversational, voice-enabled access to policy documents.

- **Offline Access**: Develop mechanisms to cache frequently accessed documents and chatbot responses locally, ensuring usability in low-connectivity environments.

- **Domain Adaptability**: Extend the chatbot framework to handle not just academic policies, but also HR manuals, corporate guidelines, healthcare protocols, and legal regulations, making it a versatile solution.

- **Security Enhancements**: Strengthen data privacy through end-to-end encryption, role-based authentication, and compliance with standards such as GDPR, ensuring sensitive academic or institutional data is safeguarded.

- **Continuous Learning Pipeline**: Automate retraining of embeddings and updating of the FAISS index when new documents are uploaded, ensuring the chatbot remains up-to-date without manual intervention.

XI. REFERENCES

[1] M. Kulkarni, P. Tangarajan, K. Kim, and A. Trivedi, *“Reinforcement Learning for Optimizing RAG for Domain Chatbots,”* in AAAI Workshop on Synergy of RL and LLMs, Seattle, USA, 2024.

[2] R. Akkiraju et al., *“FACTS About Building Retrieval Augmented Generation-based Chatbots,”* NVIDIA, 2024.

[3] C. Antico, S. Giordano, C. Koyuturk, and D. Ognibene, *“Unimib Assistant: A Student-Friendly RAG-based Chatbot,”* in AIxHMI Workshop, Bolzano, Italy, 2024.

[4] S. Neupane et al., *“Building an Informed Chatbot for University Resources,”* Mississippi State Univ., 2024.

[5] B. A. T. de Freitas and R. A. Lotufo, *“Retail-GPT: RAG for E-commerce Chat Assistants,”* in XXXII CIC UNICAMP, Brazil, 2024.

[6] D. Oreški and D. Vlahek, *“RAG in LLMs: AI Chatbot for Student Support,”* in Proc. 15th Int. Conf. e-Learning, Serbia, 2024.

[7] H. Wang et al., *“UniMS-RAG: Unified Multi-source RAG for Personalized Dialogue,”* CUHK & Univ. of Edinburgh, 2024.

[8] L. Nguyen and T. Quan, *“URAG: Hybrid RAG for University Admission Chatbots,”* HCMUT, Vietnam, 2025.

[9] U. H. Khana, M. H. Khana, *“Educational Virtual Assistant Using RAG,”* Procedia Comput. Sci., vol. 252, pp. 905–911, 2025.

[10] J. P. Nayinzira and M. Adda, *“SentimentCareBot: RAG for Mental Health Support,”* Procedia Comput. Sci., vol. 251, pp. 334–341, 2024.

[11] Jintao Liu, Ruixue Ding, Linhao Zhang, Pengjun Xie, Fie Huang (2024). *CoFE-RAG: A Comprehensive Full-chain Evaluation Framework for Retrieval-Augmented Generation with Enhanced Data Diversity*. Institute for Intelligent Computing, Alibaba Group; University of Chinese Academy of Sciences. AAAI 2024.

[12] Xiangyu Peng, Prafulla Kumar Choubey, Caiming Xiong, Chien-Sheng Wu (2025). *Unanswerability Evaluation for Retrieval-Augmented Generation*. Salesforce Research. arXiv:2412.12300.

[13] Jinyan Su, Jinpeng Zhou, Zhengxin Zhang, Preslav Nakov, Claire Cardie (2025). *Towards More Robust Retrieval-Augmented Generation: Evaluating RAG Under Adversarial Poisoning Attacks*. Cornell University; MBZUAI. arXiv:2412.16708.

[14] Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, Zizhao Zhang, Binjie Wang, Jiarong Jiang, Tong He, Zhiguo Wang, Pengfei Liu, Yue Zhang, Zheng Zhang (2024). *RAGCHECKER: A Fine-grained Framework for Diagnosing Retrieval-Augmented Generation*. Amazon AWS AI, Shanghai Jiaotong University, Westlake University. arXiv:2408.08067.

[15] Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, Zhaofeng Liu (2024). *Evaluation of*

Retrieval-Augmented Generation: A Survey. Tencent; McGill University; University of Science and Technology of China. arXiv:2405.07437.

[16] Manisha Mukherjee, Sungchul Kim, Xiang Chen, Dan Luo, Tong Yu, Tung Mai (2025). *From Documents to Dialogue: Building KG-RAG Enhanced AI Assistants*. Carnegie Mellon University; Adobe Research. FSE 2025 Companion Proceedings.

[17] Chengshuai Zhao, Riccardo De Maria, Tharindu Kumarage, Kumar Satvik Chaudhary, Garima Agrawal, Yiwen Li, Jongchan Park, Yuli Deng, Ying-Chih Chen, Huan Liu (2025). *CyberBOT: Towards Reliable Cybersecurity Education via Ontology-Grounded Retrieval Augmented Generation*. Arizona State University. arXiv:2504.00389.

[18] Jerry Huang, Siddarth Madala, Risham Sidhu, Cheng Niu, Hao Peng, Julia Hockenmaier, Tong Zhang (2025). *RAG-RL: Advancing Retrieval-Augmented Generation via RL and Curriculum Learning*. University of Illinois Urbana-Champaign; NewsBreak. arXiv:2503.12759.

[19] Arunabh Bora, Heriberto Cuayáhuitl (2024). *Systematic Analysis of Retrieval-Augmented Generation-Based LLMs for Medical Chatbot Applications*. *Machine Learning & Knowledge Extraction*, 6(4), 2355–2374. MDPI.

[20] Ahmet Yasin Aytar, Kamer Kaya, Kemal Kilic (2024). *A Retrieval-Augmented Generation Framework for Academic Literature Navigation in Data Science*. Sabanci University. arXiv:2412.15404.