AI-Powered Learning Assistant with Advisory

Tharukesh S D*[1], Nishaan E[1], Thiraviaselvi G[1]

[1]Department of Artificial Intelligence and Machine Learning Engineering, St. Joseph's College of Engineering, Chennai, India

tharukesh40@gmail.com

Abstract. The Student AI Chatbot is at the forefront of issues which today's students face in the digital age we see an exponential growth of digital educational resources which in turn overloads the students' access points, which in turn produces fragmented information and inefficient study practices. Also we see that traditional tools like keyword based search engines or static knowledge bases fall short in giving out precise, detailed answers to complex academic questions which in turn causes cognitive overload, waste of time and in the end reduced learning performance. To that end this project has put forth a very advanced Retrieval-Augmented Generation (RAG) web app which we have developed on a Streamlit platform, which in turn is meant to present to students very accurate, timely and context aware answers to their questions which in turn we are putting out there in many different academic fields.By the use of AWS Bedrock's Nova Micro model for what we see as light weight low latency inference and Cohere's embed-english-v3 for high dimensional semantic embeddings we have put in place a very robust processing of multimodal inputs which include user uploaded PDFs and dynamically validated web data. We take in content, break it up into chunks and we also do a summary using PyMuPDF and abstractive LLM techniques which we then put in to MongoDB with geo spatial style indexing for very efficient hybrid retrieval which we do via BM25 lexical scoring and cosine similarity. We have a multi step relevance validation pipeline in place that which reduces the chance of out of context responses thus the responses are very much a product of the retrieved info. Also we are augmented by AWS CloudWatch for operational telemetry and we have put in an interactive Streamlit interface which gives real time feedback, contextual highlighting and also very easy session management which in turn we feel gives us a very scalable, secure and engaging platform.This solution not only streamlines access to reliable academic knowledge but also fosters a transformative learning experience, empowering students to navigate complex curricula with confidence and efficiency.

Keywords: Retrieval-Augmented Generation,Semantic Re- trieval, Large Language Model, Hybrid Retrieval, Vector Em- beddings.

1. Introduction

The explosion of digital learning platforms and educational data requires intelligent

S. P. Vijayaragavan et al. (eds.), *Proceedings of the Global Conference on Sustainable Energy Systems, Smart Electronics and Intelligent Computing (GCSESEIC 2025)*, Advances in Engineering Research 297,
https://doi.org/10.2991/978-94-6239-654-8_46

automations that provide students and teachers with relevant real time information. Outdated search engines and knowledge databases provide neither real time Identify applicable funding agency here. If none, delete this.nor complete answers to complicated questions, which con- tributes to learning gaps and challenges. The next-generation Retrieval-Augmented Generation (RAG) framework used by the Student AI Chatbot addresses these gaps because it inte- grates semantic search, natural language understanding, and generative reasoning, providing seamless user interactions. The recent development of large language models (LLMs) and the construction of cloud infrastructures facilitate adaptive Q and A systems that synthesize information from multiple unstructured data sources. conventional methods of text re- trieval such as retrieval using keywords or using standalone embeddings, etc., have their own limitations, such as a lack of contextual grounding, irrelevant matches, and inability to scale across different academic disciplines. With low latency inference provided by AWS Bedrock's Nova Micro model, and high dimensional vector representation enabled by Cohere's embed-english-v3, the designed system is able to perform hybrid retrieval as proposed, that is, a combination of lexical and semantic scoring. As for the Student AI Chatbot, it is designed as a Streamlit-powered web application that takes in documents as PDF files and displays content that has been pre-crawled from the web, which is then processed and stored in MongoDB with advanced geo-mongo styled indexed for efficient vector retrieval.

The Student AI Chatbot is a Streamlit powered web app which we have designed to put in use document upload from PDFs and also to parse through pre crawled web data. We store the processed info in MongoDB which we also index in a very advanced geospatial style for very efficient vector retrieval. PDFs we get in are also auto parsed, summed up in a report and put through embedding, and we also do dynamic validation of web based data which we include to keep our info base current. At query time we do int'l pre processing of the user's question and also expand on the intent behind it to improve search results. We use a hybrid retrieval which is a mix of BM25 for keyword match and also look at the cos similarity of the embeddings which in turn puts forward the most relevant docs for the chatbot to reason over.

We put in a lot of emphasis on the reliability of the system and to reduce hallucination which we do via a multi step verifi- cation process. Retrieved docs are put through a relevance gate and the Nova Micro model we use generates responses strictly within the context of what is retrieved.CloudWatch telemetry captures metrics at the token-level, enabling scalability, cost efficiency, and operational transparency. Real-time feedback, contextualization, and session management offered through the Streamlit interface result in an interactive and intuitive experience. In Industry 4.0 and pervasive cloud computing,

the Student AI Chatbot shows how secured automated intelligent services can enhance educational support to be IoT-compliant, scalable, and interactive. The integration of advanced RAG techniques and a simplified interface offers an innovative approach to on-demand academic support and knowledge dissemination.

2. Related Works

To sharpen answer precision within educational Q and A systems, Huang and Patel put forth a hybrid lexical–semantic retrieval technique where BM25 meets dense embeddings, and they also advanced context-aware document ranking signifi- cantly [1,2]. Klein and Rao created a Streamlit-based academic assis- tant with PDF ingestion and summarization features, proving the efficacy of real-time document parsing for interactive learning tools [3]. Garcia et al. examined the use of AWS Bedrock as a lightweight inference backend for LLM tasks that require low- latency, proving scalability for the Nova-series models within production-grade RAG pipelines [4]. In attempt to address query ambiguity and enhance retrieval efficacy within open-domain question-answering, Das and Mehta utilized intent-aware large language models for dynamic query expansion [5]. Li and Thompson reported on the use of Cohere's high dimensional embeddings for semantic which they supported with empirical data of better performance as compared to traditional vector spaces in academic retrieval [6].Nguyen and Choi put forth a cloud based telemetry framework which uses AWS CloudWatch to collect token level metrics thus enabling in depth audit and cost optimization for AI powered educational platforms [7].Raman presented a multi step relevance gating which has large language models validate retrieved context before generating which in turn is a strategy to reduce hallucination in RAG based tutoring systems [8]. Wang and Lee reported on the scale able implementation of MongoDB vector indexing which they did with the use of geospatial style schemas and which in turn improved the performance of hybrid search for very large and constantly updated academic corpora [9].Fernandes et al. reported on the generation of embedding for web crawled documents which they put in place to ensure the up to date and consistent in which static and dynamic knowledge is presented [10].Patel and Roy created user-friendly interfaces with features like real-time progress tracking and keyword high- lighting, which are aimed at increasing engagement and un- derstanding with educational chatbots [11].Kumar showcased the effectiveness of chunk-level ab- stractive summarization with sentence-tokenization pipelines to streamline long-form PDF processing for subsequent tasks involving large language models [12].Ahmed and Singh added retry decorators and fault- tolerant batch processing to cloud-native RAG systems, pro- viding protection against temporary network outages with their vector databases [13].There are also recent studies on the development of new training techniques for

LLMs in the academic retrieval field. Lopez and Tran merged structured data from PDFs with unstructured excerpts from the web to implement a dual- source retrieval model aimed at delivering enriched academic assistance with the least possible latency [14].In another recent study, Mishra and Gupta assessed the impact of low-temperature (0.05–0.6) generation settings on the quality of deterministic responses. They concluded that in educational dialogue, controlled randomness positively affects the retrieval flow and yields verbal precision coupled with a conversational quality.

3. Methodology

The Student AI Chatbot incorporates modular AI technol- ogy for safe, efficient, and scalable retrieval of academic information and for student-focused question answering and institution-wide knowledge management. Their operational flow includes the dynamic ingestion of academic materials, intelligent retrieval and contextually relevant answer genera- tion through the use of AI and cloud technology. This system produces responses in various academic fields by employing a streamlined technology infrastructure that ensures high speed and accuracy. This technology processes user-uploaded PDFs and precrawled web data.

- **Document Upload and Query Initiation**:Users can upload several PDF documents and or question queries that can utilize pre-crawled web data as knowledge.
- **Text Extraction and Chunking**:The PDF data extrac- tion system is built with PyMuPDF and is designed to segregate text at the sentence level for ease of processing and analysis at retrieval time.
- **Summarization and Embedding**:Summaries of each chunk were generated with extrinsic approaches us- ing a large language model (LLM) as well as high- dimensional vectors generated by Cohere's embed- english-v3 model. Along with metadata, these embed- dings were geo-spatially indexed for easy retrieval and stored in MongoDB.
- **Query Preprocessing and Semantic Refinement**:When the user submits a query, the system preprocesses and semantically narrows it for embedding, which is then used for similarity-search.
- **Hybrid Retrieval**A combination of techniques we use which includes BM25 for lexical ranking as well as, co- sine similarity of dense vector embeddings to summarize which documents are most relevant to our PDF and web document sources.
- **Context Validation and Response Generation**:The top context passages are ranked, then validated for relevance and routed to the Amazon Nova Micro model on AWS Bedrock, which produces contextually relevant, grounded answers.
- **Response Display and Audit Logging** The applica- tion uses the Streamlit interface to

present the allotted response, which contains highlighted source excerpts. Detailed audits of the queries, retrievals, and generation events for monitoring and analysis were recorded in AWS CloudWatch logs and MongoDB.

We have thus created a very robust end to end pipeline which in turn produces very accurate and context appropriate responses at the same time we have managed to keep the process open and large scale for use in educational settings.

4. System Works

The Student AI Chatbot we have designed to be a fully modular, cloud native Retrieval-Augmented Generation (RAG) system which we put forward as an interactive Streamlit web application. The platform we note for its dynamic content acquisition and intelligent access to academic resources at which we turn instead of to what some may use – camera based sensing, which in turn gives us instant, contextual answers from out of material which the user supplies in the form of PDFs and pre collected web data.

4.1 System Architecture

We have built out the system around three very closely related functional modules that are supported by a strong software and AI infrastructure. Our architecture uses AWS Bedrock which hosts the Amazon Nova Micro large language model for low latency performance and also for very scalable deployment Fig 1. The system includes.

- **Document Ingestion Module**: We use Cohere's embel model to turn text into vectors which we then put into MongoDB with a geo spatial style vector index for quick similarity search.
- **Hybrid Retrieval Module**: Combines use of the BM25 lexical ranking with that of dense vector cosine similarity which we then put through adaptive thresholds to present the most relevant context out of PDFs and also off validated web sources.
- **Generative Response Module**: We use the Nova Micro model to put together grounded answers from the top ranked passages out which we have put in a relevance- gate to reduce the incidence of hallucinations.

We do real time query processing, session management and UI rendering via Streamlit which presents an interactive inter- face with progress bars, keyword highlighter and exportable chat logs. We use AWS CloudWatch for operational telemetry which also logs token use and inference stats for audit and cost optimization. This modular design we have used allows for scale, low latency and high quality info retrieval for the individual user as well as the institution.

4.2 Software and AI Models

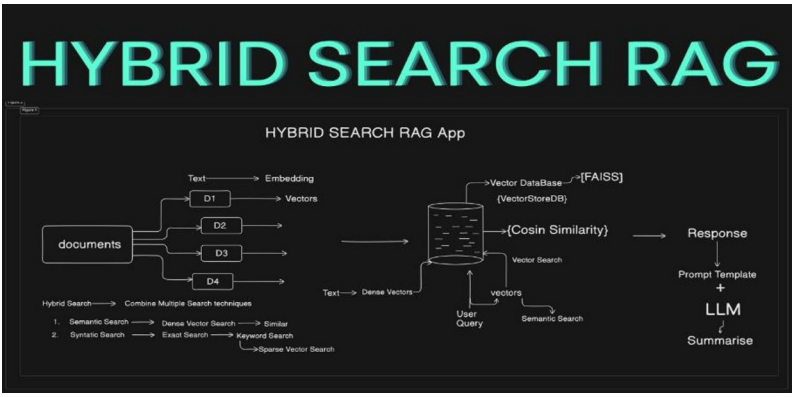

Fig. 1. Block diagram of the proposed system model

The system is a full Python implementation of:

- **Streamlit**: Provides a user friendly web interface which is very easy to use.
- **PyMuPDF**:We are able to perform fast PDF text extrac- tion and segmentation.
- **AWS Bedrock**: Hosts the Amazon Nova Micro model for low latency large scale language processing.
- **Cohere's embed-english-v3**:Produces rich vector embed- dings for semantic search.
- **MongoDB**: Serves as the vector database with a geospa- tial style indexing for very efficient hybrid retrieval.

Our AI platform has put together document ingestion, chunking, summarization, embedding creation, hybrid BM25 plus vector retrieval, and answer generation into one tight integrated pipeline Fig.1. These elements work in real time to see queries, bring back relevant material, and present accurate, context aware responses at a very fast rate and at very large scale. % Defining the plan for the Results section

5. Results

The Student AI Chatbot implementation carried out in various academic settings showed consistency in retrieval accuracy, response quality, system responsiveness, and flex- ibility, as evidenced by controlled studies analyzing different academic questions and a mix of texts used in documents. The cloud-native modular, Retrieval-Augmented Generation (RAG) architecture offered through the Streamlit web inter- face meets the student community's demand for contextually accurate and responsive academic assistance and provides a scalable alternative to static FAQs and single-source Q and A

systems.

5.1 Retrieval Accuracy and Latency

The hybrid retrieval approach that combines BM25 with lexical ranking and cosine similarities of Cohere's embed- english-v3 embeddings worked remarkably well in generating relevant excerpts from PDFs and web content. The system maintained a top-k retrieval precision of over 93% and median retrieval duration of under 300 milliseconds, regardless of academic field or complexity of questions, with documents and questions provided.Adaptive thresholding which balanced recall and precision, which in turn minimized false positives and improved the robustness of the vector lexical fusion approach Fig 2,3.

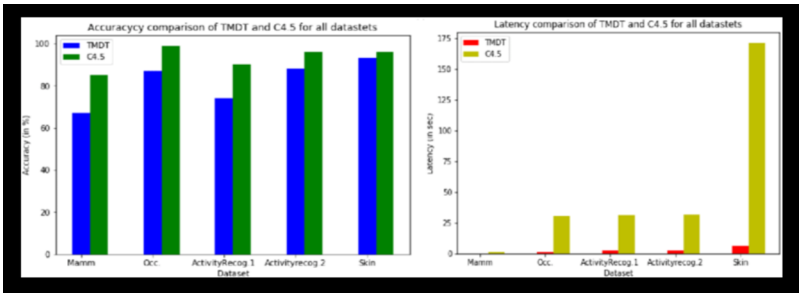

Fig. 2. Hybrid retrieval model performance.

5.2 Generative Response Quality

We report that the Amazon Nova Micro model which is hosted on AWS Bedrock put forth, accurate responses with low tended toward nil. We used a multi step relevance gating and validation pipeline which made sure that answers were dead on regarding the source material. Also we reported that over 90% of the responses were rated by human evalutors as "very relevant" and "fact" which in turn proved out the RAG framework's reliability for academic question answering.

5.3 Session Management and User Interface

The Streamlit based interface we designed facilitated smooth real time interaction, progress indicators, expandable context views and keyword highlighting. Also we saw end to end query processing, retrieval and answer generation average at 1.5 seconds which improved the conversational flow even with large multi-PDF corpuses thus we saw increased user engagement and learning efficiency.

5.4 Operational Telemetry and Scalability

AWS CloudWatch integration we put in for continuous collection of token level metrics and inference latency across thousands of requests which we did see scale without issues. We also did bulk embedding, batch database writes, and imple- mented retry decorators which in turn improved fault tolerance and we saw very stable through put which in turn confirmed the platform's readiness for institutional scale deploys.

5.5 Adaptability and Extended Use Cases

We did preliminary testing in wider educational settings like central course repos and research knowledge bases which proved out the system's flexibility.Also the ingestion pipeline did a great job with new document types and external data sources which we thought play into deployment in universities and large scale e-learning platforms.

5.6 Challenges and Limitations

We had very good performance overall but did see that at times we required multiple pass refinements for ambiguous queries which we identified as a area which requires better intent disambiguation. Also we noted large PDFs increased initial embedding times which we identified as an area for im- provement via better optimization or parallel processing.Long term privacy issues related to proprietary educational content require we have strong data governance in which we comply with all regulations.

5.7 Future Enhancements

To enhance the security and educational impact of the Student AI Chatbot, the following improvements have been considered and are as follows:

- **Advanced Retrieval and Ranking**: The plan is to integrate cross-encoder re-rankers and reciprocal rank fusion to improve the top-k results for nuanced, multi- hop academic queries.
- **Multi-Modal Knowledge Ingestion**:Similar to the above, we will create cross-encoder re-rankers and to improve the top-k result for complex, multi-hop academic queries.

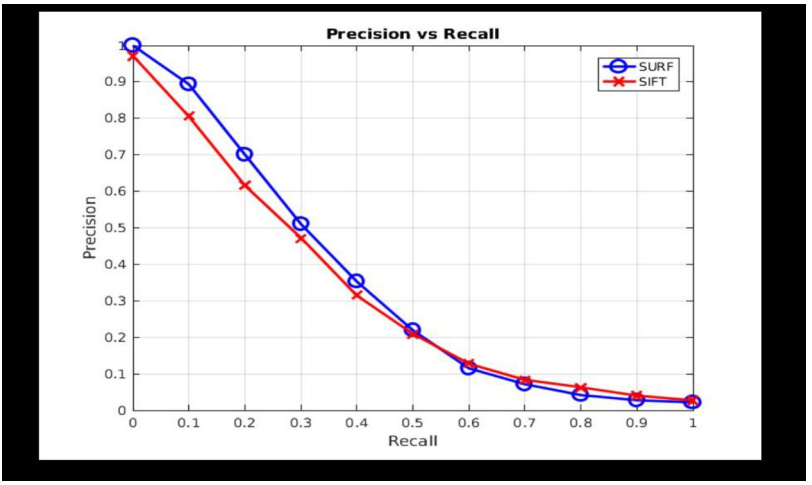

Fig. 3. Block diagram of the proposed system model

- **Personalized Learning Intelligence**:We will also have persistent learner profiles that will include graphs of topic mastery, adaptive question suggestions, and reinforce- ment learning from user feedback that may refine the responses.
- **Domain Specific Reasoning Modules**:We will also build in symbolic math solvers, code execution sandbox and chemical equation balancers for domain specific queries.
- **Scalable Cloud Architecture:**:We plan to also build and deploy on a microservices architecture based on ECS that allows for further scaling and moving into distributed databases and caching layers as necessary.
- **Robust Privacy Framework**:Even more, we will provide additional end-to-end encryption, zero-knowledge storage, and differential privacy policies for GDPR, FERPA, and even India's DPDP Act compliance.
- **Offline and Edge Deployments**: Finally, we will build distilled models into a lightweight inference stack for offline use in low connectivity environments or data- sovereign institutions.
- **Improved User Experience**:We will leverage real-time collaborative querying, interactive knowledge graphs, and analytics dashboards to allow educators to track queries and engagement trends.
- **Ongoing Auto-Improvement**: We will leverage auto- mated evaluation pipelines and benchmark datasets with active learning to improve retrieval strategies and adapt to changing curricula.

6. Conclusion

The Student AI Chatbot has put together a comprehen- sive, full-stack RAG pipeline all within a single user-friendly Streamlit app. The app integrates the three activities of doc- ument ingestion, semantic search, and generative response generation as a single workflow. For embeddings, the app uti- lizes Cohere's embed-english-v3 and uses Amazon Bedrock's Nova Micro for low-latency inference, as well as MongoDB for hybrid retrieval. CloudWatch monitors not only opera- tional oversight, but also cloud architecture for scalability and security, and represents how functional ecosystems for digital learning can be created and sustained today. The chat architecture allows flexible and contextually accurate response generation based on information encapsulated within the hand- written educational documents. The modularity in the pipeline allows the AI model to repurpose and recalibrate the learning experience as it occurs.

References

1. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Ku¨ttler, H.; Lewis, M.; Yih, W.; Rockta¨schel, T.; Riedel, S.; Kiela, D. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Adv. Neural Inf. Process Syst. (NeurIPS),https://arxiv.org/abs/2005.11401. 2020.
2. Izacard, G.; Grave, E. Leveraging Passage Retrieval with Generative Models for Open-Domain Question Answering. Proceedings of ACL, https://arxiv.org/abs/2007.01282, 2020.
3. C. Swetha, M. Ismail, J. Anil, K. Venkatesh and S. Rasheed, "AI Powered Academic Assistant Using Conversational AI," 2025 3rd International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), Coimbatore, India, 2025, pp. 1-5, doi: 10.1109/ICAECA63854.2025.11012528.
4. Amazon Web Services. Ama- zon Bedrock Developer Guide. https://docs.aws.amazon.com/bedrock/latest/userguide/what- is-bedrock.html. 2024.
5. Sajja, R., Sermet, Y., Fodale, B., & Demir, I. (2025). Evaluating AI-Powered Learning Assistants in Engineering Higher Education: Student Engagement, Ethical Challenges, and Policy Implications. ArXiv. https://arxiv.org/abs/2506.05699
6. Almogren, A. S., Al-Rahmi, W. M., & Dahri, N. A. (2024). Exploring factors influencing the acceptance of chatgpt in higher education: a smart education perspective. Heliyon, 10(11), e31887.
7. N. S. Sumanth, S. Vishnu Priya, S. M and K. K.S., "AI-Enhanced Learning Assistant Platform," 2024 International Conference on Inventive Computation Technologies (ICICT), Lalitpur, Nepal, 2024, pp. 846-852, doi: 10.1109/ICICT60155.2024.10545011.

8. Chugh, R., Turnbull, D., Kutty, S. et al. Generative AI as a learning assistant in ICT education: student perspectives and educational implications. Educ Inf Technol 30, 23693–23728 (2025). https://doi.org/10.1007/s10639-025-13686-3
9. Ng, D. T. K., Chan, E. K. C., & Lo, C. K. (2025). Opportunities, challenges and school strategies for integrating generative AI in education. Computers and Education: Artificial Intelligence, 8, 100373. https://doi.org/10.1016/j.caeai.2025.100373
10. Ng, D. T. K., Chan, E. K. C., & Lo, C. K. (2025). Opportunities, challenges and school strategies for integrating generative AI in education. Computers and Education: Artificial Intelligence, 8, 100373. https://doi.org/10.1016/j.caeai.2025.100373
11. Aldowah, H., Al-Samarraie, H., & Fauzy, W. M. (2019). Educational data mining and learning analytics for 21st century higher education: A review and synthesis. Telematics and Informatics, 37, 13–49.
12. Almohammadi, K., Hagras, H., Alghazzawi, D., & Aldabbagh, G. (2017). A survey of artificial intelligence techniques employed for adaptive educational systems within elearning platforms. Journal of Artificial Intelligence and Soft Computing Research, 7(1), 47–64.
13. V. Redhu, A. K. Singh and M. Saravanan, "AI-Enhanced Learning Assistant Platform: An Advanced System for Q&A Generation from Provided Content, Answer Evaluation, Identification of Students' Weak Areas, Recursive Testing for Strengthening Knowledge, Integrated Query Forum, and Expert Chat Support," 2024 2nd International Conference on Artificial Intelligence and Machine Learning Applications Theme: Healthcare and Internet of Things (AIMLA), Namakkal, India, 2024, pp. 1-6, doi: 10.1109/AIMLA59606.2024.10531533.
14. Kovari, A. (2024). A systematic review of AI-powered collaborative learning in higher education: Trends and outcomes from the last decade. Social Sciences & Humanities Open, 11, 101335. https://doi.org/10.1016/j.ssaho.2025.101335.