

2026

RAG Educational Chatbot Papers

DR. HAJER SALEH

AMRO ISMAIL RASHED [PAPERS 1-8]

MOHAMED ALGAZWANI [PAPERS 9-15]

Paper 1:

URAG: Implementing a Unified Hybrid RAG for Precise Answers in University Admission Chatbots – A Case Study at HCMUT

Long Nguyen and Tho Quan proposed a unified hybrid retrieval-augmented generation method, the Unified RAG (URAG) framework, to tackle the high operational costs and the risk of hallucination in university chatbots. They integrated a two-stage retrieval architecture (consisting of FAQ search and document search stages) with an in-house lightweight generator Large Language Module (LLM) to deliver precise and accurate answers to questions about university admission processes, academic programs, and campus guidelines. Instead of naive RAG or advanced RAG methods that consume huge resources, their method is to collect 300 domain-related documents from official university sources, segment them with semantic chunking, enrich the database with URAG-D (document rewrite) and URAG-F (FAQ enrichment) mechanisms, and use the dangvantuan/vietnamese-embedding model for vector representation. They compared their light-weight model (ura-hcmut/ura-llama-7b) embedded in the URAG pipeline with the top commercial models such as GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro. The systems were quantitatively assessed using an accuracy metric on a dataset consisting of 500 genuine student questions. The results showed that the URAG-powered HCMUT Chatbot achieved the best performance, outperforming the commercial models with an accuracy of 0.628, and provided a highly feasible solution for real-world academic deployments.

Paper 2:

From Questions to Insightful Answers: Building an Informed Chatbot for University Resources

Subash Neupane investigated an informed chatbot system built using Retrieval-Augmented Generation (RAG) pipelines to enhance information access and user experience for university resources. To overcome their limitations of standalone pretrained large language models (LLMs) that cannot answer domain-specific queries and hallucinate outputs, they proposed the BARKPLUG V.2 framework. Methodology: They built an external knowledge corpus by creating a multi-threaded web-scraper to collect and curate text data from 42 campus resources at Mississippi State University (MSU). The corpus was preprocessed by a recursive chunking strategy of 8000 characters and overlap of 1200 characters. Using the text-embedding-3-large model through API calls, the text was transformed to vector representations and stored in an in-memory ChromaDB. For the retrieval phase, they used a vector store backed similarity score threshold retrieval mechanism using LangChain. In the completion phase, we created accurate responses using the generator model of OpenAI, i.e., the generator model of gpt-3.5-turbo. The authors quantitatively evaluated BARKPLUG V.2 using the Retrieval Augmented Generation Assessment (RAGAS) framework on several categories, including engineering programs, general enquiries, research opportunities, and university resources. In addition, 50 students performed a subjective usability assessment using the System Usability Scale (SUS). The results showed that mean score of 0.96 and acceptance score of 67.75, indicating that the chatbot is practically feasible for real-world academic settings.

Paper 3:

From Surface Learning to Deep Understanding: A Grounded AI Tutoring System for Moodle

The authors presented the development of the AI Teaching & Learning Assistant, a modular Moodle plugin based on Retrieval-Augmented Generation (RAG) for hallucination-free education. The system's dual-centric design allows students to interact with interactive, Socratic-style tutoring, while giving educators a "Human-in-the-loop" workspace for supervised content creation. In their approach they built a presentation layer natively integrated as a PHP Moodle Activity Module using a Next.js/React frontend and a Python FastAPI backend for asynchronous parsing of PDFs, DOCX and PPTX, semantic text embedding and intent recognition. Data storage was done in a hybrid manner, using PostgreSQL, ChromaDB and MinIO. They systematically optimized the retrieval and completion phases in various disciplines, and chose the Gemini Flash 2.0 as the production model over GPT-4o-mini and Llama 3.1, as it has a higher faithfulness and speed-to-cost ratio, with an optimal retrieval context of Top-K=10. The authors quantitatively assessed the system using the RAGAS framework for STEM and Humanities disciplines and particular modes of student interaction. They also ran a usability study with 18 participants. The results showed high performance with RAGAS faithfulness scores of up to 0.97 in STEM, 0.98 in Humanities and a high user recommendation rate of 4.00 out of 5.00.

Paper 4:

Large Language Model-Based Chatbots in Higher Education

The authors investigated the integration, features, and pedagogical uses of Large Language Model (LLM) chatbots in higher education. In several generative models they contrasted various transformer architectures, training datasets, parameters and fine-tuning strategies. They discussed how LLM-based chatbots can foster highly customized instruction, provide real-time formative feedback, and support curriculum development. However, they pointed out some of the main limitations including academic dishonesty (cheating and plagiarism), data privacy breaches, algorithmic bias, and the spread of misinformation or "hallucinations". The results showed that GPT-4 is better than other models on a set of standardized human examinations. Adding an information retrieval system (IRS) to an educational question answering system improves the accuracy compared to vanilla LLM models.

Paper 5:

EduChat: A Large-Scale Language Model-based Chatbot System for Intelligent Education

The authors investigated the educational chatbot system, EduChat, based on a large-scale language model (LLM), for individualized, fair, and compassionate intelligent education. For training, they used two main datasets: (1) pre-training corpus of Chinese middle/high school textbooks, exam questions, 70k Chinese poems and 60 psychology books; (2) fine-tuning dataset of 500k customized educational instructions, including essay assessment, emotional support, Socratic teaching and open QA. Pre-processing involved semantic-level deduplication using sentence embeddings, removing one duplicate from pairs with cosine similarity above 0.7, accelerated with Compute Unified Device Architecture (CUDA) on a Graphics Processing Unit (GPU). The authors constructed structured system prompts that include personal profiles, tool usage and skill selection, and a self-check retrieval-augmented generation (RAG) mechanism. Results showed that EduChat (with retrieval) achieved the best performance on the C-Eval benchmark among other models of similar parameter scale.

Paper 6:

Reinforcement Learning for Optimizing RAG for Domain Chatbots

The authors explored a Reinforcement Learning (RL) based framework for optimizing the number of LLM tokens and reducing API costs in Retrieval-Augmented Generation (RAG) pipelines for domain-specific chatbots. They used a proprietary domain FAQ dataset of 72 FAQs relating to the credit card application process. They also used ~3.5k training, 1k validation, and 1,014 test paraphrase queries (in English and Hinglish) generated through ChatGPT and manual tagging. The pre-processing steps involved concatenating questions and answers in each FAQ to obtain vector embeddings, keeping track of conversation history by remembering the last two turns in the conversation, and formatting the inputs into specific state tokens (Classification Token [CLS] and Separator Token [SEP]) to train the external policy network. The results showed that the proposed SimThr + policy (BERT) model obtained the best performance, achieving a good balance between retrieval accuracy, generation quality, and system operating costs compared to traditional "All fetch" RAG baselines and general-purpose public embedding models.

Paper 7:

Design and Performance Evaluation of LLM-Based RAG Pipelines for Chatbot Services in International Student Admissions

Khasanova Zafar kizi, Suh investigated different pipeline configurations of LLM-based Retrieval-Augmented Generation (RAG) systems to design and evaluate an AI chatbot for international student admissions. They used an internal university admissions corpus (a 29-page PDF guidebook, 38 website FAQs and 11 department FAQs) besides two dual evaluation reference datasets: an LLM-generated QA set (72 pairs) and a human-tagged QA set (62 pairs). The pre-processing steps consisted of loading documents with metadata labels, splitting the texts using either Recursive Character TextSplitter or Semantic Chunker, converting the chunks into 384-dimensional dense vectors using the all-MiniLM-L6-v2 embedding model, and storing them in an FAISS index. The results showed that the advanced open-source OpenChat model pipeline (openchat:7b-v3.5-0106-fp16 with Recursive splitting and MultiQuery Hybrid retriever) had the best performance, outperforming the baseline commercial GPT-4o pipeline.

Paper 8:

Large Language Model based Educational Virtual Assistant using RAG Framework

Umair Hasan Khan, Muneeb Hasan Khan, Rashid Ali investigated approach to address the complexity of university systems by developing an educational virtual assistant. They combined Large Language Models (LLMs) with a Retrieval-Augmented Generation (RAG) framework to offer accurate and current information support for university admissions, course selection, and campus facilities. They used web scraping to scrape the domain-specific data from the official university website (amu.ac.in and amucontrollerexams.com) instead of the old-fashioned administrative methods. They preprocessed the collected documents (scanned files using Optical Character Recognition (OCR) tools such as pytesseract) and created the vector representations using an embedding model (all-mpnet-base-v2). They inserted the open-source models Llama-2-7b-chat-hf and Mistral-7B-Instruct-v0.2 into the RAG pipeline and tested it. The models were quantitatively evaluated based on the bilingual evaluation understudy (BLEU) score metrics. The results showed that the RAG-based Llama-2-7b-chat-hf model performed the best, surpassing the Mistral model at all n-gram levels considered and was a very feasible solution for automating administrative support.

Paper 9:

Advanced NLP Models for Technical University Information Chatbots: Development and Comparative Analysis

The authors investigated an AI chatbot for students who are seeking information about the specific courses which makes that university/institute unique. The aim of this research work was to implement a chatbot using various NLP models and compare them to identify the best one based on five chatbot models that were implemented using neural networks, TF-IDF vectorization, sequential modeling and pattern matching. Pre-Processing steps were applied included Removing Stop Word, Stemming, Lemmatization and Lancaster Stemming algorithm, which is more accurate. Furthermore, the Softmax activation function is applied to the output layer, increasing the neural network's performance. The results showed that neural network-related models had better accuracy than TF-IDF and pattern-matching models, and sequential modeling is the most accurate model because it prevents over-fitting. Furthermore, a chatbot having any kind of optimizer can improve the result, and it is most important that pattern matching and semantic analysis should be parts of a chatbot for real-time scenarios.

Paper 10:

Enhancing Arabic Information Retrieval for Question Answering

The authors investigated Arabic information retrieval for question-answering systems. They presented a comprehensive approach to addressing these issues. They developed chatbot tailored for a university community. they first created an extensive Arabic Q&A dataset by extracting data from academic documents, employing state-of-the-art Optical Character Recognition (OCR) tools. Then, they evaluated multiple text similarity measures like Pooled Fast Text Word embedding, BM25 ranking functions, and various semantic sentence embedding models. They developed web application chatbot, leveraging LangChain library and Retrieval Augmented Generation (RAG) methods, outperformed existing chatbots in domain-specific, Arabic language QA scenarios. The results showed that for Sentence Similarity: LaBSE, DistilBERT multilingual and for Context Retrieval: BM25Plus with 0.80.

Paper 11:

Improving LLM Reliability with RAG in Religious Question-Answering: MufassirQAS

The authors introduced MufassirQAS, a system that enhanced LLM accuracy and transparency using a vector database-driven Retrieval-Augmented Generation (RAG) approach. They built a dataset comprising fundamental books containing Turkish translations and interpretations of Islamic texts. This database is leveraged to answer religious inquiries while ensuring that responses remain reliable and contextually grounded. Their system presented the relevant dataset sections alongside the LLM-generated answers, reinforcing transparency. they carefully designed system prompts to prevent harmful, offensive, or disrespectful outputs, ensuring that responses align with ethical and respectful discourse. Moreover, MufassirQAS provided supplementary details, such as source page numbers and referenced articles, to enhance credibility. To evaluate its effectiveness, they tested MufassirQAS against ChatGPT with sensitive questions, and their system demonstrated superior performance in maintaining accuracy and reliability.

Paper 12:

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

Eweis et al. proposed the Retrieval-Augmented Generation (RAG) framework to improve the performance of knowledge-intensive NLP tasks by combining document retrieval with text generation. Their approach integrates a pre-trained BART sequence-to-sequence model as parametric memory with a Dense Passage Retriever (DPR) that retrieves relevant information from a large external knowledge base. The external knowledge source was built from the December 2018 Wikipedia dump, which was divided into approximately 21 million passages of around 100 words each. The framework was evaluated on several benchmark datasets, including Natural Questions (NQ), TriviaQA, WebQuestions, CuratedTrec, MS MARCO, Jeopardy Question Generation, and FEVER. The authors introduced two variants, RAG-Sequence and RAG-Token, to examine different retrieval strategies during text generation. Experimental results demonstrated that RAG-Sequence achieved the strongest overall performance, achieving Exact Match scores of 44.5% on Natural Questions and 68.0% on TriviaQA-Wiki. In addition, the RAG framework generated responses that were more accurate, factually grounded, and diverse, while reducing hallucinated information compared with a conventional sequence-to-sequence model.

Paper 13:

Student Interaction with NewtBot: An LLM-as-Tutor Chatbot for Secondary Physics Education

The authors proposed NewtBot, an LLM-based educational chatbot designed to support secondary school students in learning physics. The system was developed using GPT-3.5 and evaluated through three interaction modes: Baseline, Tutor, and Feedback, where prompt engineering was employed to enhance learning support without fine-tuning the model. The study involved 50 German secondary school students who completed eight electromagnetism tasks and post-study questionnaires measuring user experience and technology acceptance. The results showed that the Tutor mode achieved the highest user experience score ($UEQ-S = 0.985 \pm 0.439$). Furthermore, 70% of the students reported that they would use NewtBot to support their physics studies, despite 72% expressing concerns about the use of AI in education. The findings suggest that prompt-engineered LLM-based tutors can enhance students' learning experience and engagement while emphasizing the importance of addressing trust, transparency, and academic integrity issues.

Paper 14:

Centralized Database Access: Transformer Framework and LLM/Chatbot Integration-Based Hybrid Model

The authors proposed a hybrid educational chatbot model that integrates a Transformer framework with an LLM through the OpenAI API to provide efficient access to centralized educational materials. The system combines natural language processing (NLP), transformer-based text processing, and a centralized database to retrieve relevant educational content and generate context-aware responses. The implementation was developed using Python, LangChain, PyPDF2, and the OpenAI API, while lecture notes, presentations, articles, and PDF documents were used as the knowledge base. The model was evaluated using a custom centralized educational materials database instead of benchmark datasets. The results demonstrated that the proposed hybrid model improved the accessibility, organization, and retrieval of educational resources while providing accurate, context-aware responses and reducing the challenges associated with fragmented learning materials.

Paper 15:

Emerging Tensions around Learning with LLM-Based Chatbots: A CHAT Approach

The authors investigated the impact of LLM-based chatbots, such as ChatGPT, on higher education using the Cultural-Historical Activity Theory (CHAT) framework. The study explored the opportunities, concerns, and emerging tensions associated with integrating AI chatbots into learning activities. Data were collected through surveys involving 584 students and 171 teachers from a distance-learning university, and both quantitative and qualitative analyses were performed. The findings revealed that students appreciated the immediate and personalized responses provided by LLM-based chatbots, while many expressed concerns about trust, academic integrity, and the influence of AI on learning quality. The CHAT framework identified tensions related to tool mediation, division of labor, and educational rules, emphasizing the need for AI literacy, institutional policies, and collaborative guidelines to ensure the responsible integration of LLM-based chatbots into higher education.

Paper	All Models	Datasets	Dataset Link	Limitation	Best Model	Results
[1]	GPT-4o Claude 3.5 Sonnet Gemini 1.5 Pro HCMUT Chatbot	HCMUT university admissions	huggingface.co/ura-hcmut/ura-llama-7b	performance compared to larger models or third-party APIs, especially when addressing general queries beyond the admissions domain	HCMUT Chatbot	Accuracy=0.62
[2]	gpt-3.5-turbo chroma langchain	Scrapped 42 campus resources into a JSON file from the university site	Scrapped from www.msstate.edu	No Automatic Speech Recognition (ASR) capability	gpt-3.5-turbo	RAGAS score of 0.96
[3]	Gemini Flash 2.0 GPT-4o-mini Llama 3.1	teacher-provided course PDF, DOCX, and PPTX documents	tinyurl.com/mrytc2dr	Not mentioned	Gemini Flash 2.0	RAGAS score of 0.97
[4]	GPT-1, GPT-2, GPT-43, GPT-4, BERT, ROBERTa, ALBERT,	Wikipedia Toronto Book Corpus	NA	Inaccuracies generated by AI (hallucinations), misinformation, threats to data privacy,	GPT-4	Standardized Test Performance:

	ELECTRA, XLNet, T5, LaMDA	CC News		plagiarism/cheating, and dissemination of biases		- Uniform Bar Exam: 90th percentile - LSAT: 88th percentile - AP Biology, AP Macroeconomics, etc.: 5/5 - AP English Literature: 2/5
[5]	EduChat	(1) pre-training corpus of textbooks, exam questions, (2) fine-tuning dataset of 500k open QA	https://github.com/ECNU-ICALK/EduChat https://huggingface.co/ecnu-icalk	Hallucinations generation, knowledge gap in education and obsolete updates	EduChat (with Retrieval)	C-Eval Benchmark: - EduChat (with Retrieval) Avg =49.3 - EduChat (Vanilla) Avg =40.7 - Chinese-Alpaca-13B Avg =33.3
[6]	BERT GPT2 RAG	Credit Card FAQ Dataset (72 FAQs)	Not Public	Assumes the underlying RAG retriever and LLM weights are fixed. Prompt-guided evaluation is heavily dependent on GPT-4, which could introduce bias.	SimThr + policy (BERT)	Token Savings ≈31% Accuracy = 1.0
[7]	GPT-4o Ollama LLaMA3 OpenChat (latest, q6_K, fp16)	Internal University Admissions Guide (29-page PDF, FAQ1, FAQ2)	Not Public	Small size of evaluation datasets	openchat:7b-v3.5-0106-fp16	- Average RAGAS Score =0.7377 - Answer Relevancy = 0.785 - BLEU-1 = 0.5720

[8]	- Llama-2-7b-chat-hf - Mistral-7B-Instruct-v0.2 RAG	Custom university dataset built from website pdfs, scanned documents and jpegs.	Not Public	Computational constraints (reliance on Google Colab free GPU limits performance and testing).	Llama-2-7b-chat-hf (integrated with RAG)	
[9]	Sequential Modeling Neural Network model TF-IDF model	university-related information and question-answer pairs	researcher.manipal.ed	Clear communication of the chatbot's capabilities and limitations also played a crucial role in user acceptance	Sequential Modeling	Confusion matrix: Sequential model high accuracy.
[10]	BM25Plus OpenAI ada-002 LaBSE DistilBERT MPNet SBERT FastText	Arabic question-answering / information-retrieval dataset	avesis.kocaeli.edu.tr	BM25Plus it does not capture the semantics of the sentence	BM25Plus	Accuracy:0.8
[11]	ChatGPT-3.5 Turbo 16k OpenAI Embeddings LangChain Recursive Character Text Splitter TF-IDF Transformer embeddings	Quran and Hadith (Arabic)	...	Hallucination Hardware resources	MufasssirGPT	MufasssirGPT similar answers to GPT but with less hallucination
[12]	BART, DPR, RAG-Sequence, RAG-Token	Wikipedia (21M passages), Natural Questions, TriviaQA,	December 2018 Wikipedia Dump	Depends on retrieval quality and requires a large external document index	RAG-Sequence	NQ EM = 44.5, TriviaQA-Wiki EM = 68.0, improved factuality and reduced hallucinations.

		WebQuestions, CuratedTrec, MS MARCO, Jeopardy, FEVER				
[13]	GPT-3.5 Baseline, Tutor Mode, Feedback Mode	User study dataset (50 German secondary school students, 8 physics tasks)	Proprietary (Not Publicly Available)	small sample size and limited to physics education using GPT-3.5	Tutor Mode	UEQ-S = 0.985 ± 0.439 ; 70% of students would use NewtBot for schoolwork; 72% reported concerns about AI in education.
[14]	Transformer Framework, OpenAI LLM/Chatbot API, NLP, LangChain, PyPDF2	Centralized educational materials database (PDFs, presentations, articles, lecture notes)	Not Publicly Available	Evaluated on a custom educational database rather than a public benchmark dataset; performance depends on the quality of uploaded educational materials.	Transformer + LLM/Chatbot Hybrid Mode	Improved accessibility, accurate context-aware responses, efficient retrieval of educational materials, and better organization of learning resources.
[15]	CHAT Framework, LLM-based Chatbots (e.g., ChatGPT), Survey Analysis	Survey dataset (584 students and 171 teachers)	Not Publicly Available	Conducted at a single distance-learning university using survey data; findings may not generalize to other educational contexts.	CHAT Framework	LLM-based chatbots improved accessibility and personalized learning but raised concerns about trust, academic integrity, AI literacy, and educational policies .

References

- ✚ URAG: Implementing a Unified Hybrid RAG for Precise Answers in University Admission Chatbots -- A Case Study at HCMUT <https://arxiv.org/abs/2501.16276>
- ✚ From Questions to Insightful Answers: Building an Informed Chatbot for University Resources <https://arxiv.org/abs/2405.08120>
- ✚ From Surface Learning to Deep Understanding: A Grounded AI Tutoring System for Moodle <https://arxiv.org/html/2605.06963v1>
- ✚ Large Language Model-Based Chatbots in Higher Education <https://advanced.onlinelibrary.wiley.com/doi/epdf/10.1002/aisy.202400429>
- ✚ EduChat: A Large-Scale Language Model-based Chatbot System for Intelligent Education <https://arxiv.org/abs/2308.02773>
- ✚ Reinforcement Learning for Optimizing RAG for Domain Chatbots <https://arxiv.org/abs/2401.06800>
- ✚ Design and Performance Evaluation of LLM-Based RAG Pipelines for Chatbot Services in International Student Admissions <https://www.mdpi.com/2079-9292/14/15/3095>
- ✚ Large Language Model based Educational Virtual Assistant using RAG Framework <https://www.sciencedirect.com/science/article/pii/S1877050925000511>
- ✚ Advanced NLP models for Technical University Information Chatbots: Development and Comparative Analysis. *IEEE Access*, 12, 29633-29647. <https://doi.org/10.1109/ACCESS.2024.3368382>
- ✚ Enhancing Arabic Information Retrieval for Question Answering <https://avesis.kocaeli.edu.tr/yayin/330f25f8-21dc-4b50-a43d-d65a9853bcda/enhancing-arabic-information-retrieval-for-question-answering>
- ✚ Improving LLM Reliability with RAG in Religious Question-Answering: MufassirQAS <https://arxiv.org/abs/2401.15378>
- ✚ Dewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2021). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*.
- ✚ Ieb, A., & Goel, T. (2024). *Student Interaction with NewtBot: An LLM-as-Tutor Chatbot for Secondary Physics Education*. Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHIEA '24). <https://doi.org/10.1145/3613905.3647957>
- ✚ ratić, D., Šapina, M., Jurečić, D., & Žiljak Gršić, J. (2024). *Centralized Database Access: Transformer Framework and LLM/Chatbot Integration-Based Hybrid Model*. **Applied System Innovation**, 7(1), 17. <https://doi.org/10.3390/asi7010017>.
- ✚ Advanced NLP models for Technical University Information Chatbots: Development and Comparative Analysis. *IEEE Access*, 12, 29633-29647. <https://doi.org/10.1109/ACCESS.2024.3368382>