DOI: https://doi.org/10.4038/jdrra.v3i2.94

The Journal of Desk Research Review and Analysis, Vol. 3, Issue 2, 2025, 158 - 176

KNOWLEDGE RETRIEVAL REVOLUTION: LLM-BASED STUDENT Q&A SYSTEMS ENHANCED BY ONTOLOGY-DRIVEN RULES

YMBD Aberathna[1], PPGD Asanka[2] and TV Mahanama[3]

Abstract

This systematic review examines the integration of Large Language Models (LLMs) and ontology-based rule systems in developing context-aware question-answering (Q&A) platforms for academic environments, a critical advancement in addressing the challenges of personalised information retrieval from dynamic university documents. In discussing the examination of methodologies like Retrieval-Augmented Generation (RAG), vector databases (i.e., Chroma DB, Pinecone), and semantic filtering approaches, the review has documented the elements by which these technologies improve the accuracy, relevance, and adaptability of AI-generated academic assistants. The synthesis showed that a hybrid architecture, such as RAG and ontology-based filtering, dramatically increases the ability for response personalisation through enforcing batch-specific access rules and modelling hierarchical relationships within the academic domain while simultaneously reducing the hallucinations sometimes seen in outputs from generic LLMs. Ongoing challenges raised in the review include difficulties in the scale of real-time updates to ontologies, inconsistencies in the standardisation of metadata across university documents, and the lack of evaluation frameworks created for specific domains that would ultimately measure an alignment to the syllabus. The review also highlighted the need for further innovation in areas such as automated ontology generation, federated learning, which hides user activity while allowing for personalised responses, and collaborative frameworks combining personalisation with university workforces. By synthesising advancements and gaps in the field, this work provides a roadmap for developing robust, context-sensitive Q&A systems that align with the evolving needs of higher education, ultimately fostering more reliable and equitable access to academic knowledge.

Keywords: Academic Documents; LLM (Large Language Model); Ontology; RAG (Retrieval-Augmented Generation); Response Personalization

---

[1]Department of Industrial Management, University of Kelaniya, Sri Lanka.

Email: aberath-im20004@stu.kln.ac.lk https://orcid.org/0009-0007-0440-0510

[2]Department of Industrial Management, University of Kelaniya, Sri Lanka.

Email: dasanka@kln.ac.lk https://orcid.org/0000-0002-3433-5533

[3]Department of Industrial Management, University of Kelaniya, Sri Lanka.

Email: thilinim@kln.ac.lk https://orcid.org/0000-0003-0536-0040

Received date: 11.07.2025 Accepted date: 28.08.2025

Print Publishing Date: 31.10.2025 Web Publishing Date: 31.10.2025

Introduction

Universities worldwide are increasingly adopting conversational artificial intelligence systems to assist students in accessing and understanding complex academic information. Although large language models demonstrate the ability to produce natural, human-like responses, these systems frequently encounter challenges with factual precision and often fail to ensure that their answers correspond with specific academic programs or institutional guidelines. Retrieval Augmented Generation techniques have emerged as a promising solution by connecting model responses to actual course documentation and educational materials. Additionally, rule-based organisational frameworks can provide further guidance to ensure responses respect established academic hierarchies and institutional structures. This systematic review investigates how these approaches have been integrated in existing research and identifies their current limitations to establish a foundation for developing a genuinely personalised, context-sensitive chatbot explicitly designed for university environments.

Rationale

Static FAQ systems and generic AI chatbots are unable to keep up with department-specific conventions, changing curricula, and the varied backgrounds of students. Even RAG-based systems that rely on lecture notes and policy manuals treat all users in the same way, ignoring student-level information like major, batch year, and curriculum version. While academic ontologies have developed to include policy constraints and multi-level curriculum hierarchies, these formal structures are rarely incorporated into real-world chatbot implementations. Bridging these gaps promises to provide responses that are not only factually correct but also pedagogically aligned and customised to meet the needs of each learner by integrating both ontology rules and student metadata into the RAG pipeline.

Objectives

The review examines the current state of educational chatbots powered by large language models. These chatbots use advanced technologies such as RAG architectures to find and use information, ontologies to understand complex topics, and metadata-driven personalisation to tailor learning to individual students.

The main objective of this review is to identify the limitations of current chatbot systems, particularly their inability to adapt to individual students' needs and follow specific rules. Additionally, the review aims to highlight successful approaches and emerging trends in the field. Ultimately, the goal is to develop a comprehensive plan for designing and building a reliable, personalised chatbot that accurately understands educational content and earns the trust of university students.

Methodology

This review follows an improved PRISMA 2020 approach, which includes explicit protocol registration, eligibility framing, quality appraisal, and synthesis procedures, to guarantee transparency, rigour, and reproducibility.

Protocol Development And Scope

A detailed review protocol was drafted prior to screening and is provided as a supplementary document. It specifies the review's objectives, inclusion and exclusion criteria, and planned synthesis methods. The scope includes peer-reviewed journal articles and conference papers published between 2018 and 2025 that address the integration of ontology rules, RAG architectures, or student metadata personalisation in higher education chatbots.

Eligibility Criteria

Eligible studies report technical or user-centred evaluations of implementations or frameworks that integrate RAG or LLM chatbots with ontology or rule-based filtering in higher education. K–12 or corporate training environments, non-chatbot applications, non-English publications, and non-peer-reviewed sources are among the exclusion criteria. The foundational ontology papers that are allowed outside of the 2018–2025 window are listed explicitly in the protocol.

Information Sources And Search Strategy

Three primary databases, IEEE Xplore, Scopus, and Google Scholar, were systematically queried to identify relevant studies. These platforms were selected for their comprehensive coverage of technical implementations (IEEE Xplore), interdisciplinary research (Scopus), and recent conference proceedings (Google Scholar). Manual backward snowballing of reference lists and forward citation tracking were employed to supplement database searches, identifying 12 additional papers that met inclusion criteria. To mitigate platform-specific biases, search strategies were tailored to each database's syntax while maintaining conceptual consistency across queries.

Study Screening And Selection

The study selection process started with Boolean-linked keyword searches carried out across leading academic databases, including IEEE Xplore, Scopus, and Google Scholar. These searches initially resulted in 1200+ records related to large language models (LLMs), retrieval-augmented generation (RAG) systems, and ontology-based chatbots in higher education contexts. After restricting results to peer-reviewed journal articles and conference papers published between 2020 and 2025 while intentionally retaining four seminal ontology works predating 2020 till 2018 for foundational relevance, the total number was reduced to 507 items.

Following deduplication alongside initial quality control measures, such as removing non-English publications and poster abstracts, 141 records were excluded. This process resulted in 366 unique items entering the first screening.

During the first screening, titles and abstracts were evaluated against predefined inclusion criteria, focusing on studies addressing implementations or frameworks incorporating RAG, LLM-powered chatbots, and theoretical findings for ontology-driven filtering tailored explicitly for university students or higher education environments. This step removed 293 Studies mainly for targeting irrelevant domains (e.g., K–12 education, corporate training), non-chatbot AI applications, or lacking sufficient technical clarity. As a result, 73 studies proceeded to full-text review.

In the full text review, each paper was observed for empirical validation, clearly described methodologies, and direct relevance to student question answering systems at universities. This led to the removal of 43 studies—the remaining 36 studies were selected as the core of this systematic review.

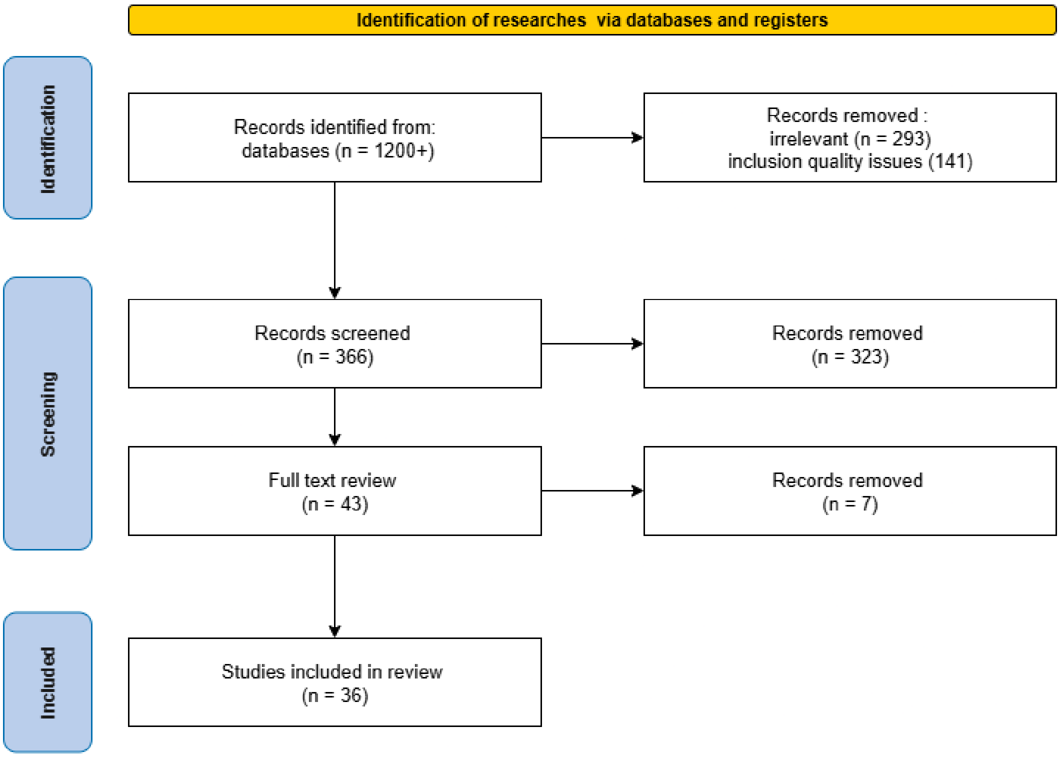

Figure 01: Literature Filtering Process

Data Collection And Quality Assessment

A standardised extraction template was developed to capture critical data elements, including study objectives, LLM architectures (e.g., GPT-4, Claude), vector database implementations (Chroma DB, Pinecone), ontology in higher education, and evaluation metrics (BLEU, RAGAS). Zotero reference management software facilitated organisation and cross-referencing of extracted data, while manual verification ensured consistency in interpreting heterogeneous reporting formats.

Synthesis Approach

The data were reviewed and organised into key themes in three steps. First, each finding was labelled with simple tags such as "metadata filtering," "ontology alignment," and "user evaluation" in order to capture its main idea. Next, these tags were sorted into three main groups: Ontology Integration in Higher Education, Retrieval-Augmented Generation in Academic Contexts, and Personalisation Mechanisms. Finally, these groups were combined into a clear framework that shows how past studies use ontologies and student information in RAG systems, highlighting common methods, differences, and emerging best practices.

Ethical And Practical Considerations

Manual screening and data extraction processes were prioritised over automation tools to preserve nuance in interpreting ontology relationships and cohort-specific filtering rules. This approach accommodated variations in terminology and ensured rigorous quality checks for studies combining semantic web technologies with generative AI actions.

Discussion Of Reviewed Literature

This section systematically reviews three interrelated sections of research published in the past years to guide the design of an ontology-aware Retrieval-Augmented Generation (RAG) chatbot for students in higher education. We start with assessing the drawbacks of general-purpose LLMs, particularly their tendency to ignore student-specific context and hallucinate. The architecture and functionality of RAG-based educational Q&A systems are then examined, highlighting how retrieval pipelines increase factual accuracy while producing off-topic results. The purpose of ontologies and rule-based filters in higher education applications is then examined, with particular attention given to metadata schemas (batch, year, department) and complex university models (e.g., Department → Course → Syllabus) that can customise chatbot responses.

Following this, we consider how student metadata and personalisation techniques have been used to adapt system behaviour, before surveying evaluation frameworks that blend information-retrieval metrics with human-centred measures. Subsequently, this led to identifying Open Gaps from the absence of ontology-integrated RAG prototypes to the absence of standards and benchmarks, which are motivating our own research aims. We, within these themes, drew on all thirty studies included in our SLR literature to map the existing landscape and expose opportunities for innovation.

Limitations Of General-Purpose LLMs In Academic Contexts

General domain LLMs such as GPT 4 and Claude are extremely linguistically talented but, when employed to undertake academic Q&A, reflect egregious deficiencies in factual basis and domain knowledge(Alkaissi & McFarlane, 2023) report a finding that 38 % of ChatGPT's responses to questions regarding medical syllabus questions contained blatant errors ranging from incorrect drug interactions to incorrectly described diagnostic criteria illustrating the danger of "artificial hallucinations" in high stakes learning environments. (Alkaissi & McFarlane, 2023; Dam et al., 2024; Maroengsit et al., 2019) Also noted that off-the-shelf chatbots do not have stringent, domain-specific assessment methodologies, so their superficial fluency usually conceals low reliability when measured against specialist academic standards. In fact, (Laato et al., 2023; Xie & Ding, 2023) both point out that foundation LLMs, which are trained on static corpora, are incapable of responding to changing curricula or institutional regulations, leading to outdated or irrelevant advice. For instance, inaccurate reporting of course requirements for a current syllabus year (Laato et al., 2023; Xie & Ding, 2023).

The overconfidence of the models worsens the problem: LLMs use a confident tone to convey outdated or hallucinogenic content, discouraging students from critically analysing it (Maryamah et al., 2024a). Despite their high surface-level coherence, Essop et al. (2023) and Ghorashi et al. (2023) show that FAQ chatbots without filtering often provide generic or contradictory answers, undermining user trust. The ability of semantic ontologies to encode complex academic structures—Curriculum → Course → Topic has long been demonstrated by (Abu-Rasheed et al., 2025; Ashour et al., 2020) and (Katis et al., 2018). However, vanilla LLMs do not take advantage of these frameworks, treating all text as if it held equal pedagogical weight. Research on open source academic chatbots (R. Joshi et al., 2024a; Wangwiwattana & Jantarick, 2024) demonstrates that responses are still out of sync with particular departmental norms or syllabus versions, even when BLEU and ROUGE scores seem high.

According to (A. T. Neumann et al., 2025), LLM deployments quickly become outdated without ongoing retraining or a document-centric base, requiring manual oversight or running the risk of spreading outdated policies. Some current studies on the use of AI chatbots in higher education indicate that while these AI tools present scalable support, they are accompanied by substantial ethical, privacy, and adoption concerns. (Yang & Evans, 2019) Observe that chatbots are likely to struggle with achieving a balance between automated assistance and sensitive guidance necessary in academic advising, describing as disadvantages the lack of adequate integration with existing learning management systems and inadequate handling of sensitive student data (Mo & Ann, 2024) promote strict ethics standards ensuring data provenance transparency, informed consent, and de-biasing, warning that without clear developer regulations, chatbots will erode student confidence and institutional reputation. Empirical evidence from (Wadhawan et al., 2023) shows that practical higher education adoption of chatbots is still patchy, with take-up hindered by a lack of confidence in accuracy, a lack of personalisation, and limited support from faculty. Meanwhile, (M. Neumann et al., 2023) caution that superficial enthusiasm for generative models like ChatGPT might be fast outpacing reflective integration and lead to over-reliance on AI-generated content and the degradation of critical thinking skills. Combined, these articles emphasise that general-purpose LLMs must be adequately regulated and contextually fashioned before being deployed into educational settings with reliability.

In conclusion, standalone LLMs are not suitable for dependable academic support without additional mechanisms for validation, ontology integration, and real-time curriculum alignment due to hallucination and overconfidence, as well as the lack of domain-aware anchoring.

RAG Architectures In Educational Q&A Systems

Retrieval-Augmented Generation (RAG) is the primary method for educational chatbots. It solves big problems of standalone LLMs, like making things up or having old information. RAG does this by saving course materials (lecture notes, syllabi, and so on) in a database you can search. When a student asks a question, RAG finds the most relevant material and gives it to an LLM (for example, GPT-4) so it can create a good answer (Benita et al., 2024; Knollmeyer et al., 2024; Maryamah et al., 2024a; Poliakov & Shvai, 2024; Wangwiwattana & Jantarick, 2024).

Early systems, like the Smart Chatbot Academic Model (Pi & Majid, 2020), set up this modular design. Later, (Maryamah et al., 2024a) used OpenAI Ada embeddings to make academic information easier to find. (A. T. Neumann et al., 2025) built MoodleBot, which matched 88% of queries exactly and had good student use. (Neupane et al., 2024) made BARKPLUG V.2, a system that scales well and works fast. (D. Joshi et al., 2024) used FAISS to build a chatbot with strong BLEU, ROUGE, and F1 scores (all above 0.8), showing how vector databases plus generative models help.

To get more accurate results, Poliakov & Shvai (2024) created Multi-Meta-RAG. This system uses an LLM to filter documents by metadata like student batch. It raised accuracy by 7% and precision by 12% on complex questions, which is very important in education. People also worked on making RAG faster and simpler. (Putri et al., 2024) cut down storage needs without lowering quality, and (Shan, 2024) made OpenRAG so schools can deploy RAG systems more easily. They focused on modular design and data privacy. These steps tackle issues of scale and upkeep in changing academic settings.

Recent studies use RAG for more student needs. (S. Chen & Yan, 2024) combined LangChain and ChatGLM with Azure AI Search to answer both academic and everyday questions, reaching ROUGE and F1 scores around 0.8 to 0.9. (Vakayil et al., 2024) tuned RAG for specific domains using Llama-2 and hit over 95% accuracy, showing how vital it is to adapt to sensitive topics. (Xie & Ding, 2023) highlighted ethical design to cut bias and mistakes in places where reliability matters most. (Mo & Ann, 2024) Also stressed responsible AI use in education.

Further work added new features. (Tjokro & Ady Sanjaya, 2024) Used knowledge graphs for traceability and provenance, boosting transparency and meeting academic standards. (Wangwiwattana & Jantarick, 2024) Created hybrid pipelines that combine FAQ extraction with LLMs to handle messy academic documents better. (R. Joshi et al., 2024a) reported that personalisation still needs work without good filtering, showing the need for detailed metadata and clear domain ontologies.

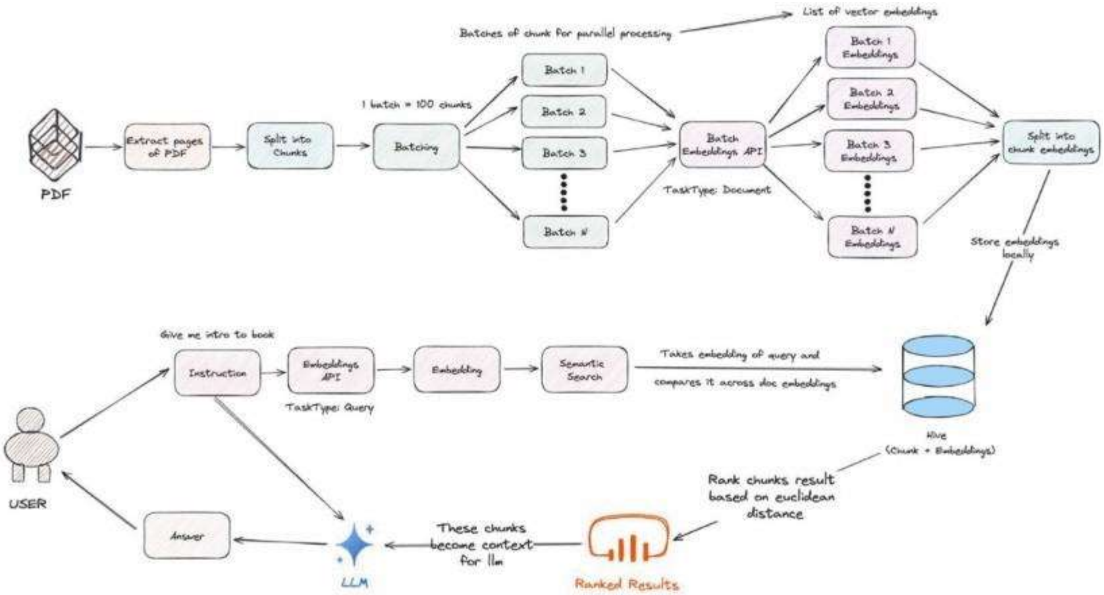

Figure 02: Generic RAG Architecture (*From (R. Joshi et al., 2024b)*)

Even though RAG clearly improves accuracy and usability, adoption is slow because of technical challenges and worries about trust (Wadhawan et al., 2023). (Mo & Ann, 2024) also call for stronger ethics, better data governance, and more openness to win stakeholder trust. Most systems still ignore student context, such as batch, year, and curriculum, which leads to generic answers. Future research, including our proposal, must mix RAG with rule-based or ontology filters to fully personalise responses by cohort and syllabus, as shown by ontology-led methods in (Katis et al., 2018), (Knollmeyer et al., 2024), and (Novogrudska & Popova, 2021).

Use Of Ontology In Higher-Education Systems

Ontology engineering gives a clear frame for modelling university structures. It defines entities like Department, Course, Staff, and Student, and shows how they relate. For example, (Ullah & Hossain, 2019) built a detailed OWL ontology with classes and subclasses and object-properties such as *teaches* and *enrolledIn*. They checked their ontology using the Fact++ and HermiT reasoners in Protégé. Their work shows that these ontologies keep complex academic data consistent and let users run semantic queries, for instance, finding all students in a given batch for a course. (Novogrudska & Popova, 2021) Also, stress how ontologies help link educational data and support dynamic semantic search in university systems.

(Katis et al., 2018) Applied this to curriculum management. They made a four-level taxonomy, Curriculum → Course → Syllabus → Topic—and added properties like *coversTopic* and *hasEvent*. They tested it using a competency-question evaluation. This modular design lets institutions reuse parts to link syllabus items and target information access. (Alrehaili et al., 2021) Extended these ideas in the HEAPAF framework by splitting Person into AcademicStaff and Student, and Course into parts like AssessmentTool and LearningOutcome. They showed how SPARQL queries and ontological rules can match faculty expertise to course modules. (Ivanova, 2022) adds to this by mapping educational ontologies to higher-level ontologies, promoting system interoperability and consistent semantics.

Full knowledge-graph deployments allow strong inferencing but can slow down real-time Q&A and raise engineering challenges. Modern RAG systems, by contrast, use vector databases for quick similarity searches instead of deep logic inference. For example, MoodleBot (A. T. Neumann et al., 2025) uses Chroma to index course PDFs, retrieves the top ten relevant chunks in milliseconds, then feeds them to GPT-4. This setup gave 88% exact-match accuracy while keeping response times acceptable. (Sharma & Sharma, 2024) found similar results: OpenAI's Ada embeddings with a cosine-similarity vector index beat more complex reranking methods in precision and recall, all under 200 ms per query.

These designs avoid the upkeep and reasoning limits of complete ontology graphs by using vector databases that scale easily and plug into LLM prompts. Ontologies remain crucial for metadata tagging, such as marking which syllabus chunks apply to which batch or department, and they form the basis for rule filters we propose to layer over vector retrieval. By combining metadata-based ontological rules with fast, scalable vector search, this hybrid method can give students accurate, context-aware answers (Knollmeyer et al., 2024; Wangwiwattana & Jantarick, 2024).

Figure 03: Complex Higher Educational Ontology (Alrehaili et al., 2021)

Student Metadata And Personalisation Techniques

To make educational chatbots more useful, responses must match student details like department, year, and batch, along with document attributes such as batch relevance and publication date. Early ontology research shows how to structure this data. (Ullah & Hossain, 2019) Modelled a university hierarchy (Student → Batch → Department) to find the right document sections. (Katis et al., 2018) improved on this by linking Curriculum → Course → Syllabus → Topic, using properties like *enrolledIn* to filter content for a specific program and year.

Recent RAG-based chatbots add metadata filtering before retrieval. (Poliakov & Shvai, 2024) Presented "Multi Meta RAG," where a small AI first picks filter criteria (for example, publication date) to limit documents before the main search. This step significantly improved answer accuracy in news tests, showing that even simple metadata helps relevance.

LLM	Accuracy		
	Ground-truth [12]	Baseline RAG [12]	Multi-Meta-RAG (ours)
GPT4 (gpt-4-0613)	0.89	*0.56*	**0.606**
PaLM (text-bison@001)	0.74	*0.47*	**0.608**

Figure 04: Accuracy Increase Due To Metadata Use(Poliakov & Shvai, 2024)

OpenRAG (Shan, 2024) takes this further by using a modular orchestration module that draws on learner preferences, interaction history, and fixed details (like department) to guide both document retrieval and how the AI answers. Still, many academic RAG systems lack strong filtering. For example, (D. Joshi et al., 2024) note that while their FAISS-based chatbot scores well on FAQs, it often gives off-topic replies for syllabus questions.

When used, metadata filters cut down information clutter, boost factual reliability, and keep students from seeing outdated or irrelevant material. However, current solutions are still piecemeal. Ontology frameworks like HEAPAF (Alrehaili et al., 2021) show promise for automating processes but have not been fully merged with RAG setups for real-time, cohort-specific filtering.
Future systems need a full academic ontology. Applying its rules during both retrieval and response generation will ensure every answer matches a student's context. This integrated model should offer adaptive help, suggesting advanced or remedial content as needed. It can also flag policy-violating queries. In doing so, chatbots will move from generic tools to personalised academic assistants, as suggested by Novogrudska & Popova (2021) and Knollmeyer et al. (2024).

Evaluation Techniques For Educational Chatbots
Evaluating educational chatbots typically focuses on two areas: retrieval quality and response quality, often using standard NLP metrics. Retrieval systems like Multi Meta RAG (A. T. Neumann et al., 2025) use ranking measures (MAP@K, MRR@K, Hit@K) to assess how well relevant information chunks are found. Answer accuracy is then checked by comparing generated responses to correct answers using overlap metrics. While these tests offer rigorous benchmarks for comparing systems, they mainly check facts and ignore whether answers fit the syllabus or are pedagogically sound.

For generated responses, conventional metrics like BLEU, ROUGE, precision, recall, and F1 are common. Studies like (Alrehaili et al., 2021) and (Jing et al., 2025) report high scores (e.g., ROUGE 0.8-0.9, F1 ~0.8), showing good fluency and content overlap. However, these metrics miss deeper context, like whether answers match a student's year or department norms. Despite this limitation, they remain widely used.

User studies provide crucial insights beyond automated scores. For example, the "myHardware" chatbot evaluation (Hidayat et al., 2022) used the System Usability Scale (SUS) with 220 students, revealing good ease-of-use but issues with complexity and consistency. Such feedback uncovers real-world strengths and weaknesses that metrics alone cannot.

A few newer studies, like (Tjokro & Ady Sanjaya, 2024), add broader measures like hallucination rates and completeness alongside automated scores and human judgments. Yet, there is still no unified framework combining content accuracy, relevance, user satisfaction, and ethics. Future work needs domain-specific benchmarks or enhanced metrics like syllabus alignment scores to ensure chatbots provide not just correct, but truly helpful and appropriate guidance.

Table 01: Categorisation of Evaluation Metrics Used.

Metric Category	Specific Metrics	Papers Using Them
Basic Retrieval Metrics	Precision, Recall, F1-Score	(Abu-Rasheed et al., 2025; S. Chen & Yan, 2024; D. Joshi et al., 2024; R. Joshi et al., 2024a; Knollmeyer et al., 2024)
NLP Overlap & Generation Metrics	BLEU, ROUGE, ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-Lsum	(D. Joshi et al., 2024; Maryamah et al., 2024a; S. Chen & Yan, 2024)
QA System Specific Metrics	Context Precision, Context Recall, Faithfulness, Answer Relevance, Faithfulness, Relevancy	(Neupane et al., 2024; Putri et al., 2024; Tjokro & Ady Sanjaya, 2024)
Information Retrieval Ranking	MAP@K, MRR@K, Hits@K	(Poliakov & Shvai, 2024)
Accuracy & Sensitivity	Accuracy, Sensitivity	(A. T. Neumann et al., 2025)
Usability & Acceptance	System Usability Scale (SUS), Technology Acceptance Model (Perceived Usefulness & Ease of Use)	(Hidayat et al., 2022; A. T. Neumann et al., 2025)
Readability & Content Quality	Flesch-Kincaid Grade Level, Amount of Information	(S. Chen & Yan, 2024)
Other Contextual Metrics	Context Relevance, Groundedness	(Wangwiwattana & Jantarick, 2024)

Research Gaps

The rapid development of large language models has significantly simplified how people find and access information across many fields, particularly in higher education. Modern language models such as GPT-4, Claude, and DeepSeek demonstrate impressive abilities to generate fluent, contextually appropriate text (Laato et al., 2023; Xie & Ding, 2023). Advanced fine-tuning techniques have enabled these models to excel at specialised tasks (D. Joshi et al., 2024). Additionally, Retrieval Augmented Generation (RAG) systems that incorporate academic documents into vector databases and ground responses in actual source materials have reduced issues with false information and outdated knowledge (Knollmeyer et al., 2024; Maryamah et al., 2024a). However, even modern enhanced systems cannot fully ensure that responses are relevant educationally or personally suitable for individual students, especially in university environments where curricula and policy guidelines change annually, and different departments follow distinct practices (Essop et al., 2023; A. T. Neumann et al., 2025).

While existing systems have achieved significant improvements in factual accuracy in recent years by up to 30% (Putri et al., 2024) and strong performance scores on standard admissions and frequently asked questions tasks (D. Joshi et al., 2024), they fundamentally treat all students the same way. If two students ask the same question, the same response will be given even though they have different academic backgrounds. These systems rarely incorporate sophisticated academic frameworks such as the comprehensive curriculum structures developed by (Katis et al., 2018) or the detailed university process models created by (Alrehaili et al., 2021). They also fail to consider how relevant academic or non-academic documents might be to individual students' backgrounds within their information retrieval processes. Some advanced systems, like Multi-Meta-RAG (Poliakov & Shvai, 2024), show how basic filtering can improve relevance. However, no academic chatbot has successfully combined these methods with formal organisational rules or real-time student information. As Pi & Majid (2020) noted, truly intelligent academic assistants need both flexible design and contextual understanding, yet the connection between theoretical frameworks and practical RAG implementation remains undeveloped.

To create genuinely personalised responses that account for each student's unique situation, including their year of study, academic department, curriculum version, and previous academic performance, future research must integrate rule-based organisational frameworks and dynamic student information throughout the entire RAG process. This integration requires three key components: first, establishing a comprehensive, university-specific organisational framework that captures curriculum relationships and institutional policies; second, collecting and updating student information in real time; and third, applying these constraints both during the initial document search phase to eliminate irrelevant content and during the response generation phase to guide appropriate answers. This comprehensive approach will maintain the natural language abilities and broad applicability of large language models while ensuring that every response is educationally sound, contextually appropriate, and customised for individual learners. Such a system will rebuild confidence in AI-powered academic tools and address the complex requirements of modern higher education.

Table 02: Comparison of RAG implementation in the literature

Paper	Evaluation Metrics & Scores
(S. Chen & Yan, 2024)	Precision: 0.9396, Recall: 0.9454, F1: 0.9425, ROUGE-1: 0.8235, ROUGE-2: 0.6875, ROUGE-L: 0.7647, Flesch-Kincaid Grade Level: 13.9, Amount of Information: 0.258
(A. T. Neumann et al., 2025)	Perceived Usefulness: 4.46, Ease of Use: 4.6, Attitude: 4.73, Behavioural Intention: 4.23, System Accessibility: 4.3, Response Accuracy: 88%, Fact-checking Accuracy: 82%
(Neupane et al., 2024)	Precision: 0.9650, Recall: 0.9750, Faithfulness: 0.9750, Relevance: 0.9750, RAGAS Harmonic Mean: 0.9675, Answer Similarity: 0.8190, Answer Correctness: 0.8626
(Poliakov & Shvai, 2024)	MRR@10: 0.6748, MAP@10: 0.3388, Hits@10: 0.9042, Hits@4: 0.792, GPT-4 Accuracy: 0.606, PaLM Accuracy: 0.608
(Putri et al., 2024)	Faithfulness: 0.8837, Answer Relevancy: 0.8998, Context Precision: 0.9831, Context Recall: 0.8675
(Wangwiwattana & Jantarick, 2024)	Context Relevance: 0.940, Answer Relevance: 0.933, Groundedness: 0.921

(Maryamah et al., 2024a)	Accuracy: 85.24%, Recall: 86%, Precision: 77%, BLEU: 25.14%, ROUGE-1: 47.24%, ROUGE-2: 34.85%, ROUGE-L: 43.34%
(R. Joshi et al., 2024a)	Accuracy: 92.5%, BLEU Score: 0.85
(Abu-Rasheed et al., 2025)	Precision: 0.9625, Recall: 0.975, F1: 0.965
(Tjokro & Ady Sanjaya, 2024)	Answer Correctness: 75.76%, Context Precision: 79.49%, Context Recall: 74%

Many recent studies have achieved strong performance across multiple evaluation metrics, showing high precision, recall, faithfulness, and relevance in chatbot responses. However, despite these improvements, most existing systems have not considered formal ontologies specific to university settings. This absence could present challenges when answering questions that require an understanding of academic hierarchies, departmental policies, or syllabus structure. Consequently, such chatbots may produce less dependable or contextually wrong responses for certain university-specific queries. However, Multi Meta RAG (Poliakov & Shvai, 2024) has demonstrated that incorporating metadata-based filtering using LLM-extracted contextual information can significantly improve retrieval quality and answer accuracy, especially for complex multi-hop questions. This shows the potential of combining ontological frameworks with metadata-based filtering to enhance personalisation and factual precision. Addressing this gap through formal ontology integration is critical for building dependable academic chatbots.

Contribution

After the systematic literature review, in the follow-up sections of this research, the following tasks will be completed. Building on these gaps, this research will develop an end-to-end RAG-based chatbot that integrates a formal, ontology-driven academic schema with live student metadata filtering. First, university ontology will be collaboratively designed, covering entities related to this research, plus rule sets for syllabus alignment and policy enforcement. Second, a metadata extraction module that ingests each user's batch, department, and syllabus version at query time, prunes the vector store accordingly, and then generates prompts that embed both the retrieved context and applicable ontology rules will be constructed. Finally, evaluation of the system against standard RAG baselines using precision/recall/F1, syllabus-alignment metrics, and a student-centred SUS study to demonstrate improved factual reliability, pedagogical relevance, and user trust, thereby filling the critical gap between ontology theory and practical RAG deployment in higher education.

Conclusion

In conclusion, this review has reviewed the progression of AI-driven chatbots in university education from general-purpose language models to advanced, retrieval-augmented systems. It has shown how foundational work on LLMs revealed both impressive fluency and concerning errors in factual accuracy and response personalisation. Subsequent innovations in RAG architectures demonstrated clear benefits by attaching responses in actual academic documents, while emerging ontology frameworks and metadata filters promise even better alignment with each student's needs.

Yet, the literature also makes clear that today's educational chatbots tend to treat all learners the same, without fully leveraging the deep structure of curricula, department-specific policies, or real-time student profiles. This gap limits their usefulness in dynamic university settings, where syllabi change each year and different and distinct requirements. Moreover, ethical considerations such as transparency about data sources and safeguards against bias remain only loosely addressed, leaving room for unintended consequences in student interactions.

Building on these discoveries, the proposed study will integrate three previously developed strands: a live metadata extraction module to encode each user's batch, department, and syllabus version; a formal academic ontology to capture curriculum hierarchies; and an RAG pipeline that utilises both components to direct generation and retrieval. By doing this, it aims to go beyond universally applicable information services and instead provide genuinely customised academic assistants who provide accurate and educationally sound responses.

Looking ahead, a crucial first step toward reliable, context-aware learning assistants is the incorporation of ontology rules and student metadata into each phase of a chatbot's workflow. Systems that combine rigorous academic grounding with human-centred design will be crucial as universities continue to adopt AI for teaching and support. Thus, this review highlights the progress made in the field as well as the obvious path forward for developing chatbots that students can depend on, semester after semester.

Descriptive Analysis Of Included Studies
Year-Wise Paper Analysis

Table 03: Year-Wise Literature Categorisation

Year	Number of Papers	Papers
2018	1	(Katis et al., 2018)
2019	3	(Maroengsit et al., 2019; Ullah & Hossain, 2019; Yang & Evans, 2019)
2020	2	(Ashour et al., 2020; Pi & Majid, 2020)
2021	2	(Alrehaili et al., 2021; Novogrudska & Popova, 2021)
2022	2	(Hidayat et al., 2022; Ivanova, 2022)
2023	7	(Alkaissi & McFarlane, 2023; Essop et al., 2023; Ghorashi et al., 2023; Laato et al., 2023; M. Neumann et al., 2023; Wadhawan et al., 2023; Xie & Ding, 2023)
2024	16	(Benita et al., 2024; S. Chen & Yan, 2024; Dam et al., 2024; D. Joshi et al., 2024; R. Joshi et al., 2024a; Knollmeyer et al., 2024; Maryamah et al., 2024b; Mo & Ann, 2024; Neupane et al., 2024; Poliakov & Shvai, 2024; Putri et al., 2024; Shan, 2024; Sharma & Sharma, 2024; Tjokro & Ady Sanjaya, 2024; Vakayil et al., 2024; Wangwiwattana & Jantarick, 2024)
2025	3	(Abu-Rasheed et al., 2025; Jing et al., 2025; A. T. Neumann et al., 2025)

Key Observations

Surge in 2024 (16 papers):

Reflects rapid advancements in LLM-based chatbots, RAG, and knowledge graphs for education in recent years.

Dominant themes: Retrieval-Augmented Generation (RAG), ethical AI, and university-specific chatbots.

2023 :

Focus on ChatGPT's impact (Alkaissi & McFarlane, 2023; Xie & Ding, 2023)

and evaluation frameworks (Essop et al., 2023; Maroengsit et al., 2019).

Pre-2020 :

Early work on ontologies (Katis et al., 2018) and rule-based systems.

2025 (2 papers):

Emerging research on vector databases (Jing et al., 2025) and LLM-driven education chatbots (A. T. Neumann et al., 2025).

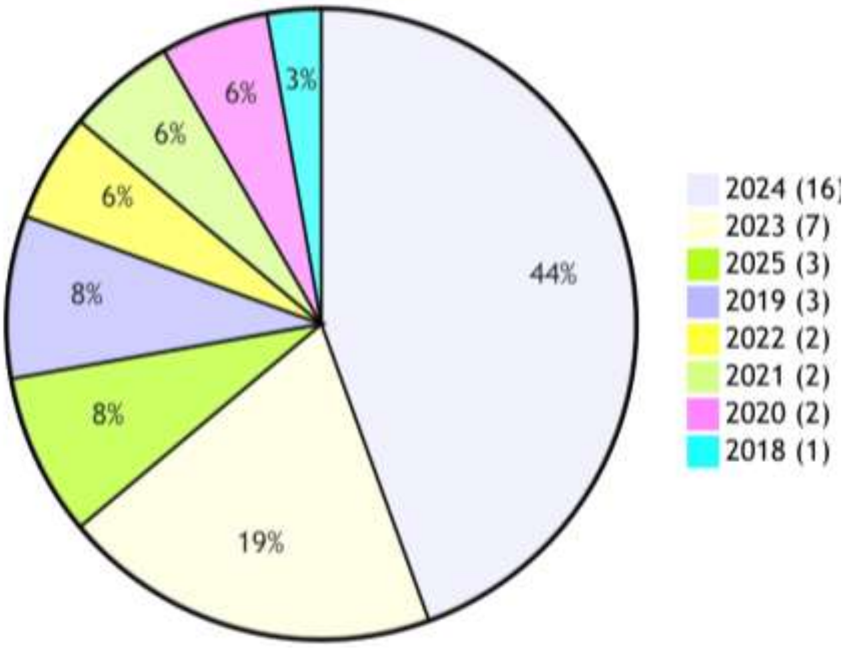

Figure 04: Year-Wise Paper Counts

Research Area Categories & Counts

Key Takeaways

- **Dominant Areas**: Higher Educational chatbots and RAG are the most researched.
- **Emerging Trends**: LLMs and ethics are gaining traction (2023–2025).

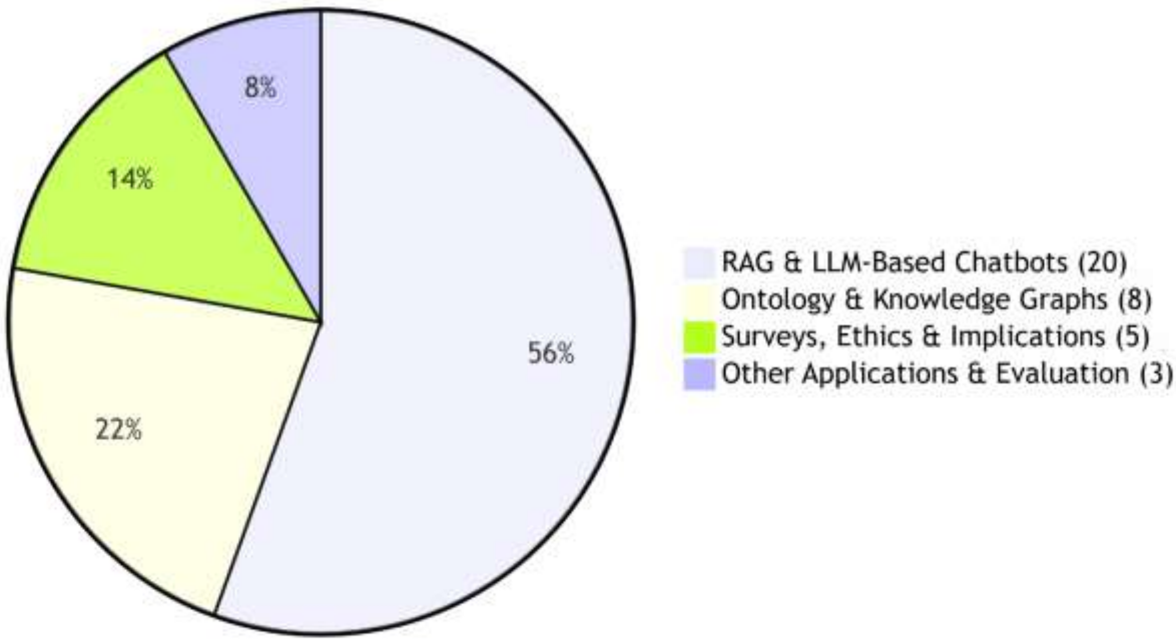

Figure 05: Research Area-Wise Count

Publisher-Wise Distribution Of Papers

Table 04: Publisher-wise Literature Categorisation

Publisher	Count	Papers
IEEE	24	(Abu-Rasheed et al., 2025; Ashour et al., 2020; Benita et al., 2024; H. Chen et al., 2024; Essop et al., 2023; Hidayat et al., 2022; Ivanova, 2022; Jing et al., 2025; D. Joshi et al., 2024; R. Joshi et al., 2024a; Knollmeyer et al., 2024; Laato et al., 2023; Mo & Ann, 2024; A. T. Neumann et al., 2025; Neupane et al., 2024; Novogrudska & Popova, 2021; Pi & Majid, 2020; Putri et al., 2024; Shan, 2024; Tjokro & Ady Sanjaya, 2024; Vakayil et al., 2024; Wadhawan et al., 2023; Wangwiwattana & Jantarick, 2024; Xie & Ding, 2023)
Springer	4	(Katis et al., 2018; Ullah & Hossain, 2019)
Cureus	2	(Alkaissi & McFarlane, 2023; Ghorashi et al., 2023)
ACM	2	(Maroengsit et al., 2019; Yang & Evans, 2019)
Wiley	1	(Alrehaili et al., 2021)
arXiv	2	(Dam et al., 2024; Poliakov & Shvai, 2024)
MDPI	1	(Sharma & Sharma, 2024)

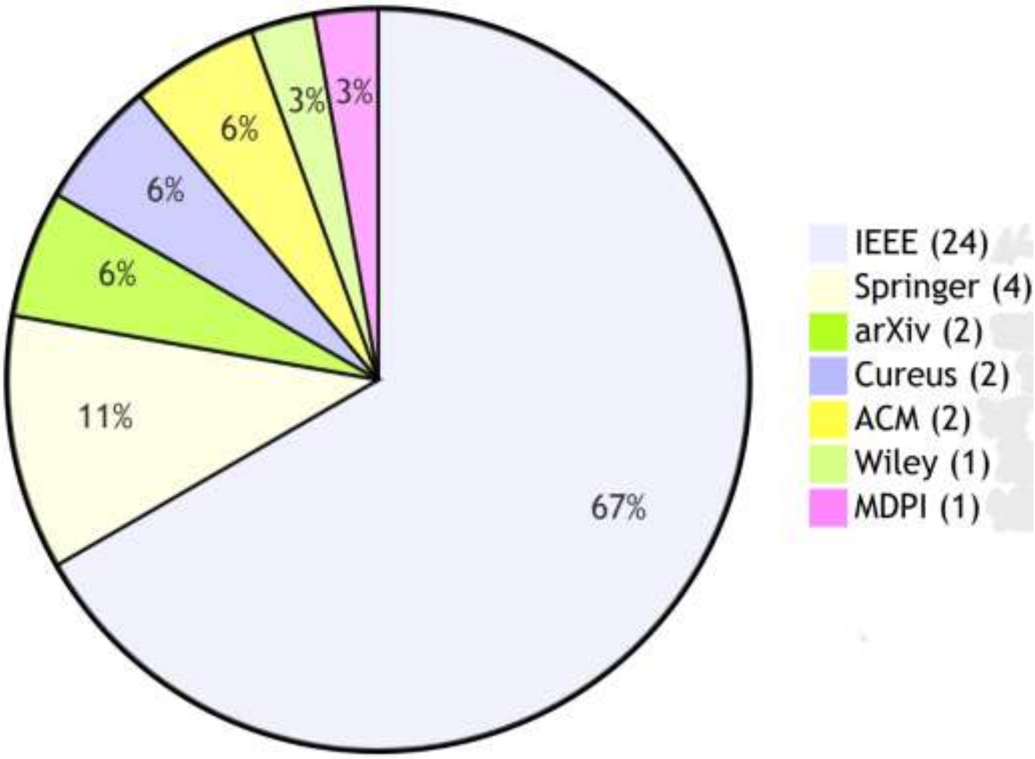

Figure 06: Publisher-wise count

Country-Wise Paper Count

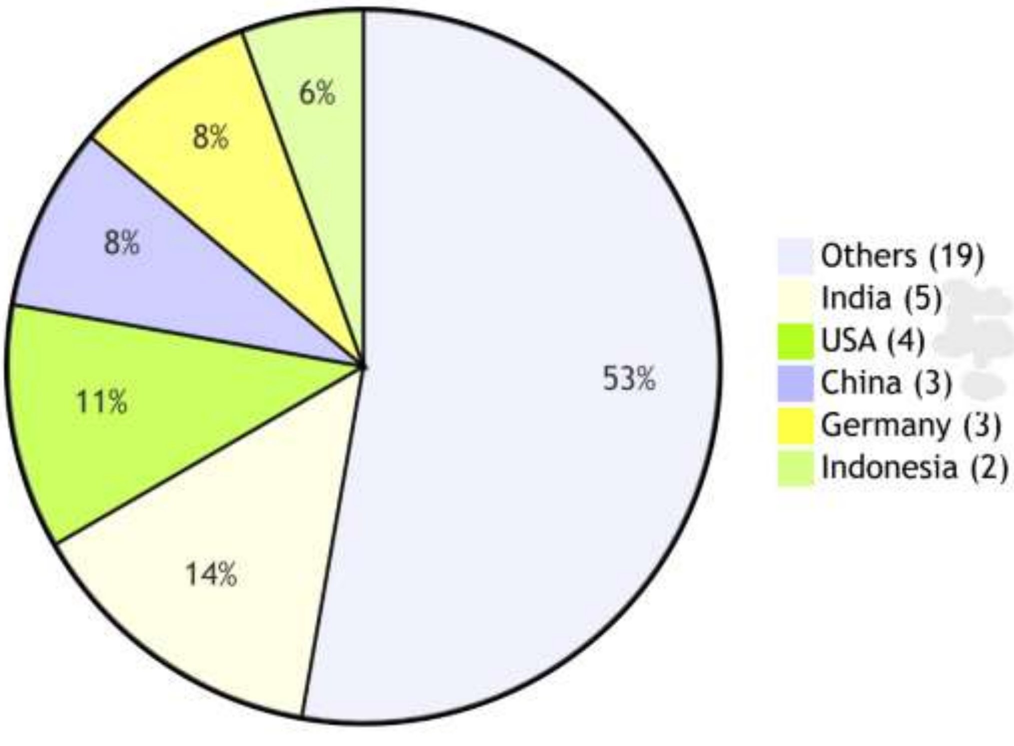

Figure 07: Country-wise paper categorisation

References

Abu-Rasheed, H., Jumbo, C., Al Amin, R., Weber, C., Wiese, V., Obermaisser, R., & Fathi, M. (2025). LLM-assisted knowledge graph completion for curriculum and domain modelling in personalised higher education recommendations. In *2025, IEEE Global Engineering Education Conference (EDUCON)* (pp. 1–5). IEEE. https://doi.org/10.1109/EDUCON62633.2025.11016377

Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: Implications in Scientific writing. *Cureus*, *15*(2), e35179. https://doi.org/10.7759/cureus.35179

Alrehaili, N. A., Aslam, M. A., Alahmadi, D. H., Alrehaili, D. A., Asif, M., & Arshad Malik, M. S. (2021). Ontology-based smart system to automate higher education activities. *Complexity*, *2021*, 5588381. https://doi.org/10.1155/2021/5588381

Ashour, G., Al-Dubai, A., Romdhani, I., & Aljohani, N. (2020). An ontological model for courses and academic profiles representation: A case study of King Abdulaziz University. In *2020 International Conference Engineering Technologies and Computer Science (EnT)* (pp. 123–126). IEEE. https://doi.org/10.1109/EnT48576.2020.00030

Benita, J., Tej, K. V. C., Kumar, E. V., Subbarao, G. V., & Venkatesh, Ch. (2024). Implementation of retrieval-augmented generation (RAG) in chatbot systems for enhanced real-time customer support in e-commerce. In *2024 3rd International Conference on Automation, Computing and Renewable Systems (ICACRS)* (pp. 1381–1388). IEEE. https://doi.org/10.1109/ICACRS62842.2024.10841586

Chen, H., Shi, N., Chen, L., & Lee, R. (2024). Enhancing educational Q&A systems using a Chaotic fuzzy logic-augmented large language model. *Frontiers in Artificial Intelligence*, *7*, 1404940. https://doi.org/10.3389/frai.2024.1404940

Chen, S., & Yan, Q. (2024). A chatbot for learning and life guide in the university based on LangChain+ChatGLM. In *2024, IEEE 4th International Conference on Software Engineering and Artificial Intelligence (SEAI)* (pp. 271–275). IEEE. https://doi.org/10.1109/SEAI62072.2024.10674054

Dam, S. K., Hong, C. S., Qiao, Y., & Zhang, C. (2024). *A complete survey on LLM-based AI chatbots* (arXiv:2406.16937). arXiv. https://doi.org/10.48550/arXiv.2406.16937

Essop, L., Singh, A., & Wing, J. (2023). Developing a comprehensive evaluation Questionnaire for university FAQ administration chatbots. In *2023 Conference on Information Communications Technology and Society (ICTAS)* (pp. 1–7). IEEE. https://doi.org/10.1109/ICTAS56421.2023.10082753

Ghorashi, N., Ismail, A., Ghosh, P., Sidawy, A., & Javan, R. (2023). AI-powered chatbots in Medical education: Potential applications and implications. *Cureus*, *15*(8), e43271. https://doi.org/10.7759/cureus.43271

Hidayat, A., Nugroho, A., & Nurfaizin, S. (2022). Usability evaluation on educational chatbot Using the system usability scale (SUS). In *2022 Seventh International Conference on Informatics and Computing (ICIC)* (pp. 1–5). IEEE. https://doi.org/10.1109/ICIC56845.2022.10006991

Ivanova, T. I. (2022). A methodology for mapping educational domain ontologies using top level ontologies. In *2022 International Conference on Information Technologies (InfoTech)* (pp. 1–4). IEEE. https://doi.org/10.1109/InfoTech55606.2022.9897119

Jing, Z., Su, Y., & Han, Y. (2025). When large language models meet vector databases: A Survey. In *2025 Conference on Artificial Intelligence x Multimedia (AIxMM)* (pp. 7–13). IEEE. https://doi.org/10.1109/aixmm62960.2025.00008

Joshi, D., Tekade, S., Zanzane, S., Sakharwade, D., Tripathi, A., & Tiwadi, S. (2024). Generative AI-driven chatbot for personalized university information and assistance. In *2024 International Conference on Artificial Intelligence and Quantum Computation-Based Sensor Application (ICAIQSA)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICAIQSA64000.2024.10882421

Joshi, R., Bubna, Y., Sahana, M., & Shruthiba, A. (2024). An approach to intelligent information extraction and utilization from diverse documents. In *2024 8th International Conference on Computational System and Information Technology for Sustainable Solutions (CSITSS)* (pp. 1–5). IEEE. https://doi.org/10.1109/CSITSS64042.2024.10816908

Katis, E., Kondylakis, H., Agathangelos, G., & Vassilakis, K. (2018). Developing an ontology for curriculum and syllabus. In A. Gangemi, A. L. Gentile, A. G. Nuzzolese, S. Rudolph, M. Maleshkova, H. Paulheim, J. Z. Pan, & M. Alam (Eds.), *The semantic web: ESWC 2018 satellite events* (Vol. 11155, pp. 55–59). Springer International Publishing. https://doi.org/10.1007/978-3-319-98192-5_11

Knollmeyer, S., Akmal, M. U., Koval, L., Asif, S., Mathias, S. G., & Groβmann, D. (2024). Document knowledge graph to enhance question answering with retrieval augmented generation. In *2024, IEEE 29th International Conference on Emerging Technologies and Factory Automation (ETFA)* (pp. 1–4). IEEE. https://doi.org/10.1109/ETFA61755.2024.10711054

Laato, S., Morschheuser, B., Hamari, J., & Björne, J. (2023). AI-assisted learning with ChatGPT and large language models: Implications for higher education. In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)* (pp. 226–230). IEEE. https://doi.org/10.1109/ICALT58122.2023.00072

Maroengsit, W., Piyakulpinyo, T., Phonyiam, K., Pongnumkul, S., Chaovalit, P., & Theeramunkong, T. (2019). A survey on evaluation methods for chatbots. In *Proceedings of the 2019 7th International Conference on Information and Education Technology* (pp. 111–119). ACM. https://doi.org/10.1145/3323771.3323824

Maryamah, M., Irfani, M. M., Tri Raharjo, E. B., Rahmi, N. A., Ghani, M., & Raharjana, I. K. (2024). Chatbots in academia: A retrieval-augmented generation approach for improved efficient information access. In *2024 16th International Conference on Knowledge and Smart Technology (KST)* (pp. 259–264). IEEE. https://doi.org/10.1109/KST61284.2024.10499652

Mo, J., & Ann, O. C. (2024). Towards responsible and ethical AI chatbot in education: A Guideline for developers. In *2024 4th International Conference on Robotics, Automation and Artificial Intelligence (RAAI)* (pp. 52–58). IEEE. https://doi.org/10.1109/RAAI64504.2024.10949557

Neumann, A. T., Yin, Y., Sowe, S., Decker, S., & Jarke, M. (2025). An LLM-driven chatbot in Higher education for databases and information systems. *IEEE Transactions on Education*, *68*(1), 103–116. https://doi.org/10.1109/TE.2024.3467912

Neumann, M., Rauschenberger, M., & Schön, E.-M. (2023). "We need to talk about ChatGPT: The future of AI and higher education. In *2023, IEEE/ACM 5th International Workshop on Software Engineering Education for the Next Generation (SEENG)* (pp. 29–32). IEEE. https://doi.org/10.1109/SEENG59157.2023.00010

Neupane, S., Hossain, E., Keith, J., Tripathi, H., Ghiasi, F., Golilarz, N. A., Amirlatifi, A., Mittal, S., & Rahimi, S. (2024). From questions to insightful answers: Building an informed chatbot for university resources. In *2024 IEEE Frontiers in Education Conference (FIE)* (pp. 1–9). IEEE. https://doi.org/10.1109/FIE61694.2024.10892994

Novogrudska, R., & Popova, M. (2021). A comprehensive review of ontology-based information systems for educational process support. In *2021, IEEE International Conference on Information and Telecommunication Technologies and Radio Electronics (UkrMiCo)* (pp. 76–79). IEEE. https://doi.org/10.1109/UkrMiCo52950.2021.9716675

Pi, S. N. M. S., & Majid, M. A. (2020). Components of smart chatbot academic model for a University website. In *2020, Emerging Technology in Computing, Communication and Electronics (ETCCE)* (pp. 1–6). IEEE. https://doi.org/10.1109/ETCCE51779.2020.9350903

Poliakov, M., & Shvai, N. (2024). *Multi-Meta-RAG: Improving RAG for multi-hop queries Using database filtering with LLM-extracted metadata* (arXiv:2406.13213). arXiv. https://doi.org/10.48550/ARXIV.2406.13213

Putri, M. R., Husodo, A. Y., & Irmawati, B. (2024). Simplification of embedding process in Retrieval augmented generation for optimizing question answering chatbot model. In *2024 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT)* (pp. 665–670). IEEE. https://doi.org/10.1109/COMNETSAT63286.2024.10862926

Shan, R. (2024). OpenRAG: Open-source retrieval-augmented generation architecture for personalized learning. In *2024 4th International Conference on Artificial Intelligence, Robotics, and Communication (ICAIRC)* (pp. 212–216). IEEE. https://doi.org/10.1109/ICAIRC64177.2024.10900069

Sharma, P., & Sharma, Y. D. (2024). Semantic pagewise augmented knowledge retrieval System: SPARKS. *Algorithms*, *17*(4), 142. https://doi.org/10.1109/ICECER62944.2024.10920365

Tjokro, V. C., & Ady Sanjaya, S. (2024). Enhancing data traceability: A knowledge graph Approach with retrieval-augmented generation. In *2024 7th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)* (pp. 473–478). IEEE. https://doi.org/10.1109/ISRITI64779.2024.10963652

Ullah, M. A., & Hossain, S. A. (2019). Ontology-based information retrieval system for University: Methods and reasoning. In A. Abraham, P. Dutta, J. K. Mandal, A. Bhattacharya, & S. Dutta (Eds.), *Emerging technologies in data mining and information security* (Vol. 814, pp. 119–128). Springer Singapore. https://doi.org/10.1007/978-981-13-1501-5_10

Vakayil, S., Juliet, D. S., Anitha, J., & Vakayil, S. (2024). RAG-based LLM chatbot using Llama-2. In *2024 7th International Conference on Devices, Circuits and Systems (ICDCS)* (pp. 1–5). IEEE. https://doi.org/10.1109/ICDCS59278.2024.10561020

Wadhawan, I., Jain, T., & Galhotra, B. (2023). Usage and adoption of chatbot in education Sector. In *2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 1097–1103). IEEE. https://doi.org/10.1109/ICICCS56967.2023.10142511

Wangwiwattana, C., & Jantarick, W. (2024). Building intelligent academic service agents: FAQ extraction and chatbots with large language models. In *2024 8th International Conference on Information Technology (InCIT)* (pp. 370–375). IEEE. https://doi.org/10.1109/InCIT63192.2024.10810553

Xie, X., & Ding, S. (2023). Opportunities, challenges, strategies, and reforms for ChatGPT in Higher education. In the *2023 International Conference on Educational Knowledge and Informatisation (EKI)* (pp. 14–18). IEEE. https://doi.org/10.1109/EKI61071.2023.00010

Yang, S., & Evans, C. (2019). Opportunities and challenges in using AI chatbots in higher Education. In *Proceedings of the 2019 3rd International Conference on Education and E-Learning* (pp. 79–83). ACM. https://doi.org/10.1145/3371647.3371659