

# Comparative Study of Knowledge Graph-Augmented and Retrieval-Augmented Large Language Models for Educational Chatbots: A Review

Archna Pandey<sup>1</sup>, Mrs. Arifa Khan<sup>2</sup>

<sup>1</sup>Master of Technology, Computer Science and Engineering, Lucknow Institute of Technology, Lucknow, India

<sup>2</sup>Assistant Professor, Department of Computer Science and Engineering, Lucknow Institute of Technology, Lucknow, India

\*\*\*

**Abstract** - Large Language Models (LLMs) have emerged as powerful engines for building conversational educational agents, yet their tendency to hallucinate, their reliance on static parametric knowledge, and their limited support for multi-hop pedagogical reasoning restrict their direct deployment in classrooms and e-learning platforms. Two complementary augmentation paradigms have been proposed to address these limitations: Retrieval-Augmented Generation (RAG), which grounds LLM outputs in externally retrieved unstructured or semi-structured text, and Knowledge Graph (KG)-augmented LLMs, which ground generation in structured, relationally connected facts. This paper presents a systematic, comparative review of these two paradigms as applied specifically to educational chatbots. We first describe the architectural foundations of RAG and KG-augmented LLMs, then examine hybrid GraphRAG approaches that combine both. We compare the two paradigms along eight dimensions relevant to education: factual accuracy, reasoning depth, curriculum alignment, personalization, explainability, latency, construction cost, and scalability supported by evidence drawn from recent literature. Representative educational systems built on each paradigm are surveyed, including syllabus assistants, course-knowledge-graph tutors, and liberal-arts collaborative chatbots. The review finds that RAG offers faster deployment and broader topical coverage with lower engineering overhead, while KG-augmented LLMs offer superior curriculum traceability, multi-hop concept reasoning, and personalized learning-path generation, at the cost of graph-construction effort. We conclude that hybrid KG + RAG architectures represent the most promising direction for next-generation educational chatbots and outline open challenges and future research directions for M.Tech. and doctoral researchers entering this field.

**Key Words:** Large Language Models, Retrieval-Augmented Generation, Knowledge Graphs, GraphRAG, Educational Chatbots, Intelligent Tutoring Systems, Natural Language Processing, Personalized Learning.

## 1. INTRODUCTION

Conversational artificial intelligence has moved rapidly from rule-based dialogue trees to neural sequence models and, most recently, to Large Language Models (LLMs) such as GPT-4 and its successors, which are trained on web-scale corpora and exhibit strong general-purpose language understanding and generation abilities [28]. This capability

has generated substantial enthusiasm for LLM-based educational chatbots that can answer student questions, explain concepts, generate practice problems, and provide round-the-clock tutoring support, functions traditionally associated with Intelligent Tutoring Systems (ITS) and computer-assisted instruction. Empirical reviews of LLM adoption in higher education document rapid uptake across chatbots, learning-content generation, automated assessment, and tutoring tools, alongside emerging evidence of benefits and risks [12], [13].

Despite this promise, LLMs deployed 'as-is' in educational settings face three interlocking problems. First, LLMs encode knowledge implicitly in their parameters and are known to hallucinate producing fluent but factually incorrect statements which is unacceptable in a pedagogical context where incorrect explanations can entrench student misconceptions [24]. Second, an LLM's parametric knowledge is frozen at the point of training and cannot reflect an institution's specific syllabus, textbook, or the most recent curriculum revision, a serious limitation for course-specific assistants [14], [21]. Third, general-purpose LLMs have no explicit model of a learner's mastery of prerequisite concepts, and therefore cannot easily support personalized, prerequisite-aware tutoring without external scaffolding [22], [48].

Two broad augmentation strategies have been proposed in the wider NLP literature to mitigate these problems, and both have since been adapted to education. Retrieval-Augmented Generation (RAG), introduced by Lewis et al. [1], couples a parametric LLM with a non-parametric retriever that fetches relevant passages from an external corpus (e.g., lecture notes, textbooks, FAQ archives) at inference time, allowing the model to condition its answer on retrieved evidence. RAG has since been extended into naive, advanced, and modular variants with increasingly sophisticated pre-retrieval, retrieval, and post-retrieval mechanisms [2], [3]. Knowledge Graph (KG)-augmented LLMs instead ground generation in an explicit, structured graph of entities and relations for instance, a course-concept graph in which nodes represent topics such as 'Binary Search Tree' or 'AVL Rotation' and edges represent prerequisite or is-a relations enabling multi-hop, relationally explicit reasoning and transparent provenance for every fact used in an answer [4], [7].

These two paradigms are frequently discussed independently in the technical literature, yet from an educational-technology standpoint they represent two distinct philosophies for knowledge grounding: RAG treats a course's knowledge as a searchable document collection, whereas KG-augmentation treats it as a navigable concept map. A systematic, side-by-side comparison specifically oriented toward educational chatbot design choices factual grounding, curriculum traceability, personalization, reasoning depth, and deployment cost has been comparatively underexplored, motivating the present review.

This paper makes the following contributions. (i) It synthesizes the architectural principles of RAG and KG-augmented LLMs relevant to conversational educational agents. (ii) It reviews representative educational chatbot systems built on each paradigm, as well as hybrid GraphRAG systems that combine both. (iii) It develops a structured, multi-dimensional comparison framework and applies it to contrast the two paradigms. (iv) It identifies open challenges and outlines future research directions suitable for M.Tech. and early doctoral research. The remainder of this paper is organized as follows. Section II presents background concepts. Section III describes the review methodology. Sections IV and V review RAG-based and KG-augmented educational chatbots, respectively. Section VI discusses hybrid GraphRAG approaches. Section VII presents the comparative analysis. Section VIII surveys applications and case studies. Section IX discusses challenges, Section X outlines future directions, and Section XI concludes the paper.

## 2. BACKGROUND AND FOUNDATIONAL CONCEPTS

### 2.1. Large Language Models

LLMs are transformer-based neural networks trained on massive text corpora using self-supervised objectives, most commonly next-token prediction. Scaling model size, data, and compute has produced emergent capabilities such as few-shot in-context learning, instruction following, and multi-step reasoning [28]. However, all knowledge in a vanilla LLM is stored implicitly in its weights (parametric memory), which makes the model's knowledge static, opaque, and difficult to update or audit properties that are particularly problematic in education, where content must remain aligned with an evolving, verifiable curriculum.

### 2.2. Retrieval-Augmented Generation (RAG)

RAG augments an LLM with a non-parametric memory: an external corpus indexed for similarity search, typically using dense vector embeddings and approximate nearest-neighbour search. At inference time, a query is embedded, the most similar passages are retrieved, and these passages are concatenated with the query as additional context for the

LLM's generation step [1]. Gao et al. [2] categorize the evolution of RAG into three generations: Naive RAG (index–retrieve–generate), Advanced RAG (which adds pre-retrieval query rewriting/expansion and post-retrieval re-ranking or compression), and Modular RAG (which allows flexible, task-specific arrangements of retrieval, routing, and generation modules, including iterative and agentic retrieval loops) [10]. Huang and Huang [3] similarly decompose the RAG pipeline into pre-retrieval, retrieval, post-retrieval, and generation stages, each offering optimization opportunities such as query transformation, hybrid sparse–dense retrieval, contextual compression [9], and re-ranking. RAG's central appeal is that it decouples 'what the model knows' from 'what the model has memorized,' allowing knowledge to be updated simply by editing the retrieval corpus rather than retraining the model.

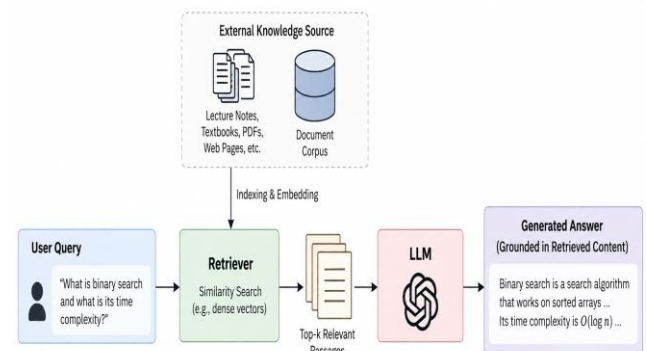


**Figure 1: Retrieval-Augmented Generation (RAG)**

### 2.3. Knowledge Graphs and KG-Augmented LLMs

A Knowledge Graph is a structured representation of knowledge as a set of triples (head entity, relation, tail entity), forming a labelled, directed multigraph. In an educational KG, nodes typically represent concepts, learning objectives, or curriculum units, and edges represent semantic relations such as prerequisite-of, part-of, or related-to [23], [52]. Pan et al. [4] formalize three general frameworks for unifying LLMs and KGs: (1) KG-enhanced LLMs, in which KG triples or subgraphs are injected into the LLM during pre-training, fine-tuning, or inference (e.g., via prompt-based fact injection) to improve factual grounding and interpretability; (2) LLM-augmented KGs, in which the LLM assists KG construction, completion, and embedding tasks; and (3) Synergized LLMs + KGs, in which the two representations interact bidirectionally, with the LLM performing reasoning over graph-retrieved subgraphs and the KG providing structured, traceable evidence for generated answers. KG-augmentation is particularly attractive in education because a graph makes explicit the prerequisite structure of a subject, enabling multi-hop reasoning (e.g., linking a student's question about 'eigenvalues' back through 'characteristic polynomial' and

'determinants') that flat document retrieval cannot easily express [7], [48], [49].

## 2.4. GraphRAG: A Convergence of the Two Paradigms

A recent line of work, popularized by Microsoft's GraphRAG [6] and surveyed comprehensively by Han et al. [5], treats the knowledge graph itself as the retrieval index for a RAG pipeline. An LLM is first used to extract entities and relations from a document corpus, building a graph; community-detection algorithms then group related entities, and LLM-generated community summaries are pre-computed. At query time, retrieval operates over graph neighbourhoods, paths, or community summaries rather than flat text chunks, allowing the system to answer both local, fact-specific queries and global, corpus-wide synthesis queries [5], [6]. This hybridization directly motivates the comparative and complementary treatment adopted in this review.

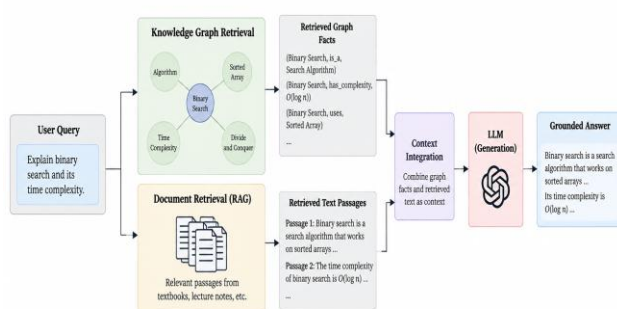


**Figure 2: GraphRAG: A Convergence of the Two Paradigms**

## 3. REVIEW METHODOLOGY

This review follows a structured, narrative-synthesis methodology suitable for a technical review paper rather than a formal Cochrane-style systematic review with meta-analysis, given the heterogeneity of system architectures and evaluation metrics reported in the source literature. Candidate literature was identified from arXiv, IEEE Xplore, ACM Digital Library, ScienceDirect, and Springer, using search strings combining ('retrieval-augmented generation' OR 'RAG') AND ('large language model'), ('knowledge graph' AND 'large language model') AND ('education' OR 'tutoring' OR 'chatbot'), and ('GraphRAG' OR 'graph-based retrieval'). Papers published primarily between 2020 and 2026 were prioritized, reflecting the timeline of LLM-era RAG and KG-augmentation research, while foundational pre-LLM references on knowledge-graph-based educational question answering were retained where they establish design patterns still used today [21], [49].

Inclusion criteria were: (i) the work proposes or evaluates a RAG or KG-augmented LLM architecture; (ii) the work is either explicitly applied to an educational/tutoring use case, or is a general architectural/survey paper providing foundational technical grounding cited by educational applications. Exclusion criteria removed papers that discuss LLMs in education without any retrieval or knowledge-graph augmentation component, and non-peer-reviewed blog content except where used solely to describe an industry system (e.g., Microsoft GraphRAG) also documented in an accompanying technical report. Screened works were organized into four thematic buckets: (a) RAG architecture and optimization, (b) KG-LLM unification, (c) GraphRAG/hybrid methods, and (d) educational chatbot/ITS applications, which structure Sections IV–VIII below. A comparison framework of eight evaluation dimensions (Section VII) was then derived inductively from recurring themes factual accuracy/hallucination, reasoning depth, curriculum alignment, personalization, explainability, latency, construction/maintenance cost, and scalability across the surveyed papers.

## 4. RETRIEVAL-AUGMENTED LLMs FOR EDUCATIONAL CHATBOTS

The dominant motivation for applying RAG to educational chatbots is domain and curriculum grounding: an institution can build a chatbot that answers strictly from its own syllabi, lecture slides, and textbooks without fine-tuning a foundation model. Cooper and Klymkowsky [21] argue that RAG-based chatbots let instructors evaluate and support student reasoning by grounding generative answers in vetted course material rather than the model's unconstrained parametric knowledge. Thway et al. [21] report a classroom study of a RAG chatbot's impact on learning in a chemistry course, while Thüs et al. [21] describe a RAG system designed to increase student engagement with scientific literature both illustrating the common pattern of indexing course-specific documents (slides, papers, readings) as the retrieval corpus.

A representative architecture, described in the context of a 'local RAG-based Intelligent Tutoring System for domain-specific education,' ingests curriculum documents, chunks and embeds them, stores them in a vector index, and at query time retrieves the top-k relevant chunks to condition the LLM's response, with the explicit design goal of curriculum-grounded, reproducible educational AI evaluable through classroom trials [21]. Kuhail et al. [15] conducted a systematic review of educational chatbots more broadly and found that retrieval and knowledge-base grounding are recurring design requirements for chatbots intended to answer subject-matter questions reliably, predating but foreshadowing LLM-era RAG designs.

Within this pipeline, several engineering choices materially affect educational usefulness. Chunking strategy determines

whether a retrieved passage preserves enough context to answer a question correctly; naive fixed-length chunking can split a worked example from its explanatory text, degrading answer quality, whereas semantically aware chunking that respects section or slide boundaries tends to preserve pedagogical coherence [3], [9]. Embedding-model choice affects whether domain-specific terminology, such as discipline-specific notation in mathematics or engineering, is correctly matched to relevant passages, motivating the use of domain-adapted or fine-tuned embeddings in several of the surveyed systems [26], [32]. Re-ranking and contextual compression stages, discussed by Verma [9], are particularly relevant in education because course corpora often contain many superficially similar passages, such as repeated definitions across different weeks of a course, and naive top-k retrieval can therefore surface redundant or lower-quality material ahead of the most pedagogically appropriate explanation.

Post-retrieval generation strategy is equally important. Several educational RAG systems deliberately avoid simply concatenating retrieved text into the answer, instead prompting the LLM to synthesize, simplify, or Socratically reframe the retrieved material to match a student's apparent level of understanding, echoing findings from neural dialog-tutoring research on the pedagogical value of guided questioning over direct answer delivery [24]. This distinguishes educational RAG from generic document-QA RAG: the objective is not merely to answer correctly but to answer in a way that supports learning, which introduces additional evaluation criteria such as clarity, scaffolding, and appropriate difficulty, beyond retrieval precision and answer factuality alone.

Beyond single-document QA, RAM2C demonstrates a retrieval-augmented multi-role, multi-expert collaboration architecture for a liberal-arts educational chatbot, in which multiple specialized retrieval-grounded 'expert' agents (e.g., a pedagogy expert and a subject-matter expert) collaboratively draft and refine a response before it is delivered to the student, aiming to balance factual accuracy with pedagogically appropriate scaffolding rather than simply dumping retrieved facts [16]. This exemplifies the Modular RAG paradigm [2] applied to a tutoring context, where retrieval is only one module among several coordinated components (planning, role-play, verification).

A separate application, the 'Internet Wizard,' extends the RAG idea to open-domain educational question answering by using live web search including search-engine knowledge-graph snippets as the retrieval source rather than a fixed local corpus, allowing the chatbot's knowledge base to remain current for fast-changing or broad-coverage topics; a pilot study in higher education reported high user satisfaction with the resulting answers [20]. Complementing document- and web-level retrieval, 'Platform-Independent and Curriculum-Oriented Intelligent Assistant for Higher Education' targets logistics and policy questions (grading,

deadlines, academic integrity) by retrieving from a structured syllabus knowledge model, illustrating that RAG corpora in education range from unstructured lecture text to semi-structured syllabus templates [22].

Architecturally, most educational RAG systems adopt the standard retrieve-then-generate pipeline with course-specific adaptations: (1) domain-restricted corpora (syllabi, slides, past exam banks) rather than web-scale indices; (2) stricter grounding/verification steps, since factual errors have direct pedagogical consequences [21]; and (3) additional post-retrieval pedagogical shaping, such as Socratic reformulation of retrieved facts into guiding questions rather than direct answers, as advocated in dialog-tutoring research [24]. The principal strength of the RAG paradigm for education is rapid deployability: an institution can stand up a working, reasonably grounded chatbot by indexing existing course materials, without the substantial upfront engineering cost of constructing a formal knowledge graph.

## **5. KNOWLEDGE GRAPH-AUGMENTED LLMS FOR EDUCATIONAL CHATBOTS**

KG-augmented educational chatbots take a structurally different approach: rather than retrieving loosely related text passages, they retrieve or reason over an explicit concept graph that encodes the relational structure of a subject. A systematic literature review of knowledge-graph construction and application in education identifies five recurring application domains: adaptive and personalized learning, curriculum design and planning, concept mapping and visualization, semantic search and question answering, and miscellaneous applications, concluding that KGs consistently support tailored, data-driven educational experiences across these domains [23], [52].

A concrete pre-LLM-era example is the course-knowledge-graph question-answering system, which constructs a Chinese university-course knowledge graph and builds an intelligent QA system on top of it specifically to compensate for the difficulty of timely teacher-student communication in online education and to support personalized knowledge retrieval [49]. Although this system predates modern LLM integration, it establishes the graph-construction and graph-querying pattern entity extraction from course materials, relation labelling (e.g., prerequisite, sub-topic), and graph-based query answering that LLM-era KG-augmented chatbots subsequently combine with generative fluency.

The 'Personalized Learning Tutor' concept illustrates how a student-specific knowledge graph, tracking which concepts a learner has mastered versus which remain weak, can be consulted by an LLM before it generates an explanation, quiz, or worked example, so that the system 'fills gaps rather than repeating things already mastered' and can explicitly state prerequisite relations such as Math → Algebra → Quadratic

Equations → Factorization [48]. This pattern separating a static domain KG (the subject's concept map) from a dynamic learner KG (an individual's mastery state) recurs across the personalized-learning literature and is central to why KG-augmentation is often favoured over plain RAG for adaptive tutoring: the graph structure makes prerequisite gaps and next-best-concept recommendations explicit and computable (e.g., via graph traversal or shortest-path queries) rather than implicit in retrieved text similarity [23], [50].

At the architectural level, Pan et al.'s KG-enhanced LLM framework [4] and the reliable-KG framework proposed by Zhang [7] describe how graph triples or retrieved subgraphs are converted into natural-language or structured prompts that are injected into the LLM's context, analogous to RAG's passage injection but with the crucial difference that the injected unit is a verified relational fact or reasoning path rather than an unstructured passage, which improves interpretability because each asserted fact can be traced to a specific triple and provenance source in the graph [7]. This traceability directly supports auditability requirements in education, where instructors may need to verify why a chatbot recommended a particular remediation path.

Graph construction methodology also varies across the surveyed systems. Some rely on manual or semi-manual curation by subject-matter experts, which yields high-precision graphs at the cost of scalability, while others use LLM-assisted entity and relation extraction from existing course text, trading some precision for substantially faster construction, an approach consistent with the 'LLM-augmented KG' direction described by Pan et al. [4]. Regardless of construction method, graph quality assurance typically involves cross-checking extracted relations against a reference syllabus or textbook table of contents, and iteratively correcting missing or erroneous edges using structure-based error-detection techniques such as those proposed by Zhang [7], which contrast candidate triples against surrounding graph structure to flag likely inconsistencies before they reach students.

KG-augmentation, however, carries a substantially higher upfront cost: constructing a reliable, sufficiently complete course knowledge graph requires curriculum-specific entity and relation extraction (often semi-automated, with instructor validation), ongoing maintenance as syllabi evolve, and error-correction pipelines, since noisy or incomplete graphs can propagate incorrect prerequisite links or missing relations into generated tutoring content [7]. This construction and maintenance burden is the principal trade-off against the interpretability and personalization gains discussed above, and is examined further in the comparative analysis in Section VII.

## **6. HYBRID APPROACHES: GRAPH RAG FOR EDUCATION**

Because RAG and KG-augmentation address complementary weaknesses RAG offers broad, low-cost topical coverage while KG-augmentation offers structured, multi-hop reasoning and traceability a growing body of work combines them under the umbrella of GraphRAG. Han et al. [5] survey this convergence and note that, unlike conventional RAG where retriever, generator, and corpus are uniformly represented in embedding space, graph-structured retrieval must additionally handle heterogeneous, domain-specific relational schemas, which is a non-trivial design challenge. Microsoft's GraphRAG constructs an LLM-derived entity graph from a document corpus and precomputes hierarchical community summaries, enabling the system to answer both narrow factual queries (via local graph neighbourhoods) and broad, corpus-level synthesis queries (via community summaries), addressing a known weakness of flat-chunk RAG on 'global' questions that require aggregating information scattered across many documents [5], [6].

Lightweight variants such as LightRAG introduce a two-stage retrieval pipeline balancing coverage and efficiency, while HippoRAG combines a knowledge graph with personalized PageRank to propagate relevance from query-seed entities across multiple hops in a single retrieval pass, and GFM-RAG applies graph neural networks to strengthen multi-hop reasoning further [5]. Domain-adapted variants such as SemRAG integrate semantic chunking with graph-guided retrieval and generation stages specifically to improve grounding for question-answering tasks [32], and graph-enhanced RAG has also been demonstrated for structured customer-support domains, combining subgraphs from a domain KG with text passages retrieved from a support archive to produce more coherent, factually grounded answers a pattern directly transferable to a course-support chatbot that must reconcile a syllabus KG with free-text lecture notes [25].

For educational chatbots specifically, a GraphRAG-style architecture would typically layer (i) a domain concept graph capturing prerequisite and topical relations, (ii) a document-level vector index for full lecture/textbook text, and (iii) an orchestration layer that decides, per query, whether to traverse the graph (for 'why do I need to know X before Y' or multi-step reasoning queries), retrieve text passages (for 'explain this paragraph' or definitional queries), or combine both (for open-ended tutoring dialogue). This is consistent with the multi-role, multi-expert collaboration pattern already observed in RAM2C [16] and with the modular RAG taxonomy's routing mechanisms [2], [10], suggesting that hybrid architectures are a natural extension rather than a radical departure from existing educational-chatbot engineering practice.

## 7. COMPARATIVE ANALYSIS

This section compares RAG and KG-augmented LLMs across eight dimensions distilled from the literature reviewed in Sections IV–VI, with particular attention to their implications for educational chatbot design. Table I summarizes the comparison; a narrative discussion of each dimension follows.

**TABLE I: Comparison of RAG-Based and KG-Augmented Educational Chatbots**

| Dimension                           | RAG-Based Educational Chatbots                                                          | KG-Augmented Educational Chatbots                                                          |
|-------------------------------------|-----------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------|
| Knowledge unit retrieved            | Unstructured/semi-structured text chunks (passages, slides, FAQs) [1]–[3]               | Structured triples, entities, and relational subgraphs [4], [7]                            |
| Factual accuracy / hallucination    | Improved vs. vanilla LLM, but can still mis-synthesize across retrieved passages [24]   | Higher precision on relation-specific facts; errors traceable to specific triples [4], [7] |
| Multi-hop / prerequisite reasoning  | Limited; similarity search does not natively encode prerequisite chains [3]             | Strong; graph traversal explicitly encodes prerequisite/sub-topic paths [48], [49]         |
| Curriculum alignment & traceability | Depends on corpus curation; provenance is a source document, not a fact-level link [21] | High; every asserted fact/edge is traceable to a curriculum-validated triple [7], [23]     |
| Personalization to learner state    | Requires external session memory; not natively concept-aware                            | Native; learner-mastery sub-graph can be intersected with domain KG [48], [50]             |
| Explainability                      | Passage-level citation only                                                             | Fact- and path-level explanation (which prerequisite led to which conclusion) [4]          |

| Dimension                       | RAG-Based Educational Chatbots                                                    | KG-Augmented Educational Chatbots                                                                            |
|---------------------------------|-----------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------|
| Construction & maintenance cost | Low-moderate: chunk and embed existing documents [2]                              | High: entity/relation extraction, validation, ongoing graph maintenance [7], [23]                            |
| Latency & scalability           | Efficient dense vector search scales well to large corpora [3]                    | Graph traversal can be costly at scale; community-summary precomputation mitigates this in GraphRAG [5], [6] |
| Coverage breadth                | High — any indexed document is queryable, including open web sources [20]         | Bounded by graph completeness; incomplete graphs create knowledge gaps [7]                                   |
| Typical educational use case    | Syllabus/FAQ Q&A, literature-engagement chatbots, open-domain tutoring [20], [21] | Concept-map tutoring, personalized learning paths, curriculum planning [23], [48], [49]                      |

### 7.1. Factual Accuracy and Hallucination Mitigation

Both paradigms reduce hallucination relative to a vanilla LLM by grounding generation in external evidence, but they fail differently. RAG systems can hallucinate by miscombining facts from multiple retrieved but topically adjacent passages, or by retrieving passages that are semantically similar to the query yet factually irrelevant [3], [24]. KG-augmented systems instead fail primarily through graph incompleteness or noise an absent or incorrect edge silently removes a valid reasoning path or introduces an invalid one but when a fact is present in the graph, its correctness is typically higher because it has been extracted and (ideally) validated as a discrete relational assertion rather than inferred from loosely related prose [4], [7].

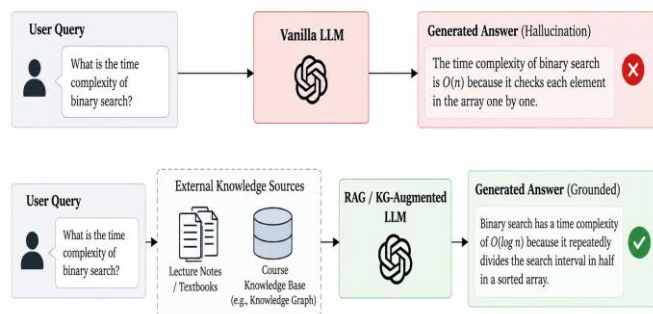


**Figure 3: Factual Accuracy and Hallucination Mitigation**

## 7.2. Reasoning Depth and Personalization

The clearest advantage of KG-augmentation for education is native support for multi-hop, prerequisite-aware reasoning and personalization: because the domain graph explicitly encodes concept dependencies, a system can compute a learner's shortest remediation path (e.g., 'to understand eigenvalues, first review determinants, then characteristic polynomials') by graph traversal rather than by hoping that semantically similar passages happen to cover the same dependency chain [4], [48], [49]. RAG systems can approximate this through careful prompt engineering or iterative/agentive retrieval loops [10], but such approximations are heuristic rather than structurally guaranteed.

## 7.3. Deployment Cost and Practicality

Conversely, RAG's principal advantage is low barrier to entry: an institution with existing digitized course materials can build a reasonably grounded chatbot in a matter of days by chunking and embedding documents [2], [21], whereas building a validated course knowledge graph is a substantially larger undertaking requiring domain-expert-in-the-loop entity and relation extraction and ongoing maintenance as curricula evolve [7], [23]. This cost asymmetry likely explains why the majority of deployed educational chatbots surveyed in this review (Section IV) use RAG or RAG-like retrieval, while KG-augmented systems remain comparatively more common in research prototypes and specialized personalized-learning platforms (Section V).

## 7.4. Convergence via GraphRAG

The comparative analysis above suggests that neither paradigm strictly dominates the other for education: RAG wins on coverage breadth, deployment speed, and scalability, while KG-augmentation wins on reasoning depth, traceability, and personalization. This complementarity is precisely what motivates the hybrid GraphRAG architectures

discussed in Section VI, and is consistent with Pan et al.'s 'Synergized LLMs + KGs' framework, which treats structured and unstructured knowledge sources as jointly, rather than competitively, exploited [4]–[6].

## 8. APPLICATIONS AND CASE STUDIES IN EDUCATION

This section synthesizes representative deployed or evaluated systems discussed in the reviewed literature, organized by paradigm, to ground the comparative claims of Section VII in concrete evidence.

### 8.1. RAG-Based Case Studies

- Domain-specific ITS: a local RAG pipeline indexing course documents to support reproducible, curriculum-grounded tutoring, with an explicit mixed-methods evaluation protocol for classroom trials [21].
- RAM2C: a liberal-arts chatbot combining retrieval with multi-role, multi-expert collaborative response generation to balance factual grounding with pedagogical framing [16].
- Internet Wizard: an open-domain educational QA chatbot that augments retrieval with live web search and search-engine knowledge-graph snippets, evaluated via a higher-education pilot study with favourable user satisfaction [20].
- Syllabus assistant: a curriculum-oriented assistant retrieving from a structured syllabus knowledge model to answer course-logistics questions [22].
- Chemistry-education RAG chatbots evaluated for their effect on student reasoning and engagement with scientific literature [21].

### 8.2. KG-Augmented Case Studies

- Course knowledge-graph QA system: a Chinese university-course KG paired with an intelligent QA system to support personalized knowledge retrieval and compensate for limited teacher availability in online learning [49].
- Personalized Learning Tutor: an architecture pairing a domain concept KG with a per-student mastery KG to drive gap-focused explanation and quiz generation [48].
- Educational KG systematic review findings: recurring use of KGs for adaptive/personalized learning, curriculum design, concept mapping/visualization, and semantic QA across the surveyed literature [23], [52].

### 8.3. Cross-Domain Evidence Informing Educational Design

Evidence from adjacent domains further substantiates the comparative claims above. In biomedicine and healthcare, KG-augmented LLM surveys report that structured grounding meaningfully improves complex, multi-hop question-answering performance and reduces hallucination relative to ungrounded generation, reinforcing the reasoning-depth advantage attributed to KG-augmentation in Section VII [4], [17]. In e-commerce customer support, graph-enhanced RAG combining domain-KG subgraphs with retrieved support documents produced more coherent, factually grounded responses than either component alone, empirically supporting the hybrid GraphRAG argument of Section VI [25]. Surveys of medical-teacher usage of LLM chatbots in India similarly highlight both the appeal and the accuracy concerns of ungrounded LLMs in educational contexts, motivating the grounding strategies reviewed in this paper [19].

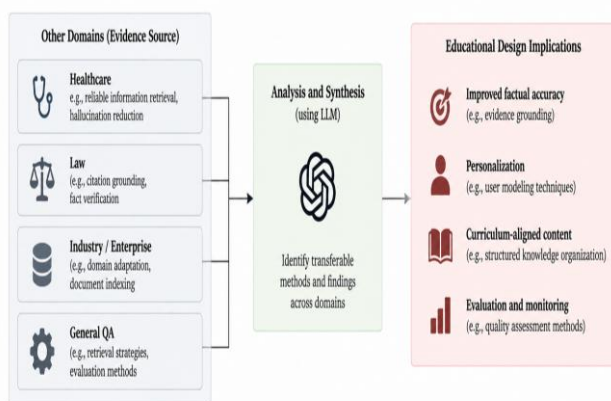


**Figure 4: Cross-Domain Evidence Informing Educational Design**

## 9. CHALLENGES AND OPEN ISSUES

### 9.1. Evaluation Standardization

Reviewed systems use heterogeneous evaluation protocols user-satisfaction surveys [20], classroom mixed-methods trials [21], or benchmark QA accuracy making direct, quantitative comparison between RAG- and KG-augmented educational chatbots difficult. There is no widely adopted, education-specific benchmark analogous to general-purpose RAG evaluation suites, hampering rigorous comparative research.

### 9.2. Knowledge Graph Construction and Maintenance

Constructing a sufficiently complete and accurate course knowledge graph remains labour-intensive, and

automatically extracted graphs are prone to noise, incompleteness, and drift as curricula evolve, requiring error-detection and completion pipelines such as those proposed by Zhang [7]. This is a major practical barrier to the wider adoption of KG-augmented educational chatbots relative to the comparatively lightweight corpus-indexing required for RAG.

### 9.3. Hallucination and Trust

Even with retrieval or graph grounding, LLMs can misattribute, over-generalize, or fabricate connective reasoning between retrieved facts [24]. In education, where students may not be equipped to critically evaluate an authoritative-sounding but incorrect explanation, this residual hallucination risk carries particular pedagogical weight and motivates continued research into fact verification and citation-level grounding.

### 9.4. Personalization at Scale

Maintaining an individualized learner-mastery knowledge graph for every student in a large course, and keeping it synchronized with assessment data, raises scalability and privacy questions that are only beginning to be addressed in the literature [48], [50].

### 9.5. Latency and Infrastructure Cost

Graph traversal and community-summary precomputation in GraphRAG-style systems introduce additional indexing and query-time costs relative to simple vector search, a trade-off that institutions with limited technical infrastructure must weigh against the reasoning-quality benefits [5], [6].

### 9.6. Equity and Accessibility

Most reviewed systems are demonstrated on well-resourced institutional corpora or high-resource languages; extending both RAG and KG-augmentation to low-resource languages, under-digitized curricula, and low-bandwidth deployment contexts remains comparatively underexplored in the surveyed literature.

## 10. FUTURE RESEARCH DIRECTIONS

- Standardized educational benchmarks: development of shared, curriculum-grounded QA and tutoring-dialogue benchmarks that allow like-for-like comparison of RAG, KG-augmented, and hybrid GraphRAG educational chatbots.
- Automated, instructor-in-the-loop KG construction: semi-automated pipelines that extract candidate concept graphs from existing course materials using LLMs, with lightweight instructor validation, to

reduce the construction-cost barrier identified in Section IX.

- Adaptive hybrid orchestration: learned routers that decide, per query type, whether to invoke document retrieval, graph traversal, or both, extending the modular RAG and GraphRAG patterns already emerging in the literature [2], [5], [10].
- Fine-grained, fact-level explainability: techniques that expose not just source documents but specific triples or reasoning paths used in generating a tutoring response, building on KG-based interpretability work [4], [7].
- Privacy-preserving learner-state modelling: federated or on-device approaches to maintaining individual learner-mastery graphs without centralizing sensitive student data [48].
- Multilingual and low-resource educational KG-RAG systems: extending both paradigms to under-served languages and institutions, addressing the equity gap noted in Section IX.
- Longitudinal classroom evaluation: moving beyond short pilot studies toward semester- or year-long controlled trials that measure actual learning-outcome effects of RAG versus KG-augmented tutoring, extending early efforts such as [21].

## 11. CONCLUSION

This review has compared Retrieval-Augmented Generation and Knowledge Graph-augmented Large Language Models as two complementary strategies for building grounded, trustworthy educational chatbots. RAG-based systems offer rapid deployability, broad topical coverage, and scalable retrieval, making them well-suited to syllabus and FAQ-style assistants that can be built directly from existing course documents. KG-augmented systems, in contrast, offer explicit multi-hop reasoning, fact-level traceability, and native support for prerequisite-aware personalization, at the cost of substantially higher graph-construction and maintenance effort. Neither paradigm is strictly superior across all educational use cases; rather, the evidence surveyed here indicates that they address different failure modes of ungrounded LLMs and are increasingly being combined in hybrid GraphRAG architectures that route between text retrieval and graph reasoning depending on query type. For M.Tech. and doctoral researchers entering this space, the most promising avenues lie in standardized evaluation, cost-efficient semi-automated knowledge-graph construction, adaptive hybrid orchestration, and rigorous, longitudinal classroom evaluation of learning outcomes directions that would move educational chatbot research from isolated proof-of-concept systems toward pedagogically validated, deployable technology.

## REFERENCES

- [1] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 9459–9474.
- [2] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, and H. Wang, "Retrieval-augmented generation for large language models: A survey," *arXiv:2312.10997*, 2023.
- [3] Y. Huang and J. X. Huang, "A survey on retrieval-augmented text generation for large language models," *arXiv:2404.10981*, 2024.
- [4] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu, "Unifying large language models and knowledge graphs: A roadmap," *arXiv:2306.08302*, 2023.
- [5] H. Han et al., "Retrieval-augmented generation with graphs (GraphRAG)," *arXiv:2501.00309*, 2025.
- [6] Microsoft Research, "GraphRAG: Unlocking LLM discovery on narrative private data," *Microsoft Research Blog*, 2024. [Online]. Available: <https://www.microsoft.com/en-us/research/blog/graphrag-unlocking-llm-discovery-on-narrative-private-data/>
- [7] Q. Zhang, "Enhancing large language models with reliable knowledge graphs," *arXiv:2506.13178*, 2025.
- [8] L. Mei, S. Mo, Z. Yang, and C. Chen, "A survey on multimodal retrieval-augmented generation," *arXiv:2504.08748*, 2025.
- [9] S. Verma, "Contextual compression in retrieval-augmented generation for large language models: A survey," *arXiv:2409.13385*, 2024.
- [10] "Agentic retrieval-augmented generation: A survey on agentic RAG," *arXiv:2501.09136*, 2025.
- [11] "Retrieval-augmented generation for natural language processing: A survey," *arXiv:2407.13193*, 2024.
- [12] D. Yigci, M. Eryilmaz, A. K. Yetisen, S. Tasoglu, and A. Ozcan, "Large language model-based chatbots in higher education," *Adv. Intell. Syst.*, vol. 7, no. 2400429, 2025.
- [13] "Large language models in education: A systematic review of empirical applications, benefits, and challenges," *Computers and Education: Artificial Intelligence*, 2025.

- [14] H. Xu, W. Gan, Z. Qi, J. Wu, and P. S. Yu, "Large language models for education: A survey," arXiv:2405.13001, 2024.
- [15] M. A. Kuhail, N. Alturki, S. Alramlawi, and K. Alhejori, "Interacting with educational chatbots: A systematic review," *Educ. Inf. Technol.*, vol. 28, no. 1, pp. 973–1018, 2023.
- [16] "RAM2C: A liberal arts educational chatbot based on retrieval-augmented multi-role multi-expert collaboration," arXiv:2409.15461, 2024.
- [17] "LLM agents for education: Advances and applications," arXiv:2503.11733, 2025.
- [18] "Designing a local RAG-based intelligent tutoring system for domain-specific education," in *Proc. Springer Conf. Ser.*, 2025.
- [19] A. Deb Roy, I. B. Jaiswal, D. N. Tiu, D. Das, S. Mondal, J. K. Behera, and H. Mondal, "Assessing the utilization of large language model chatbots for educational purposes by medical teachers: A nationwide survey from India," *Cureus*, vol. 16, no. 11, e73484, 2024.
- [20] "Internet wizard for enhancing open-domain question-answering chatbot knowledge base in education," *Applied Sciences*, vol. 13, no. 14, p. 8114, 2023.
- [21] M. M. Cooper and M. W. Klymkowsky, "Let us not squander the affordances of LLMs for the sake of expedience: Using retrieval augmented generative AI chatbots to support and evaluate student reasoning," *J. Chem. Educ.*, vol. 101, no. 11, pp. 4847–4856, 2024.
- [22] "Platform-independent and curriculum-oriented intelligent assistant for higher education," arXiv:2302.09294, 2023.
- [23] "A systematic literature review of knowledge graph construction and application in education," *Heliyon*, 2024.
- [24] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Comput. Surv.*, vol. 55, no. 12, Article 248, 2023.
- [25] P. Patel, "Graph-enhanced retrieval-augmented question answering for e-commerce customer support," arXiv:2509.14267, 2025.
- [26] "SemRAG: Semantic knowledge-augmented RAG for improved question-answering," arXiv:2507.21110, 2025.
- [27] "A survey on unifying large language models and knowledge graphs for biomedicine and healthcare," in *Proc. 31st ACM SIGKDD Conf. Knowledge Discovery and Data Mining (KDD)*, 2025.
- [28] OpenAI, "GPT-4 technical report," arXiv:2303.08774, 2023.
- [29] L. Huang et al., "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Trans. Inf. Syst.*, vol. 43, no. 2, pp. 1–55, 2025.
- [30] "Intelligent question and answer system based on course knowledge graph," in *Proc. IEEE Conf.*, 2023.